IMPROMPTU TIMED-WRITING AND PROCESS-BASED TIMED-WRITING EXAMS: COMPARING STUDENTS' PERFORMANCE AND INVESTIGATING STUDENTS' AND RATERS' PERCEPTIONS

By

Virginia David

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Second Language Studies – Doctor of Philosophy

2015

ABSTRACT

IMPROMPTU TIMED-WRITING AND PROCESS-BASED TIMED-WRITING EXAMS: COMPARING STUDENTS' PERFORMANCE AND INVESTIGATING STUDENTS' AND RATERS' PERCEPTIONS

By

Virginia David

In this study I compare 81 students' performances on and perceptions of two different writing exams: an impromptu timed-writing (TW) exam and a process-based timed-writing (PBTW) exam. The students had 45 minutes to write an essay for the impromptu TW exam. For the PBTW exam, the same participants read an article and watched two short videos about a topic, discussed the topic in small groups, and planned their essays. This took approximately 45 minutes. After that, they had 45 minutes to write their essays. Thus, the PBTW exam lasted 90 minutes. After taking both exams, the participants answered a short post-writing questionnaire about their perceptions of the two exams. Eighteen participants were randomly selected or volunteered to participate in a semi-structured interview in groups to receive more detailed information about their opinions regarding the exams. My secondary aim was to investigate what raters think of the two exams. Two raters scored the essays using an analytic rubric and participated in a semi-structured interview after they scored the essays. Furthermore, the raters participated in two training and norming sessions before they began rating the essays, both of which were audio-recorded to gather more information about the raters' perceptions of the exams.

To explore the results, I correlated the scores that the students received in the exams in SPSS and performed a *t*-test to determine whether there were significant differences between the scores. I also examined the essays to investigate accuracy, lexical and syntactic complexity, and

fluency. I investigated the students' perceptions of the exams by analyzing their answers in the post-writing questionnaire and the transcripts of the semi-structured interviews. Finally, I analyzed the training and norming sessions, as well as the semi-structured interviews with the raters to explore their perceptions of the exams.

The results of the study revealed that, although the learners' performance did not significantly differ in the two exams, the participants expressed a clear preference for the PBTW exam because they had time to learn about the topic through the article, videos, and discussion, they had time to plan their writing, and they could use the ideas in the source materials to support their opinions in the essays. The scores that the test takers received for content and punctuation were significantly higher in the PBTW exam, while the scores that they received for spelling were significantly higher in the TW exam. The participants also wrote significantly longer essays and significantly more words per minute in the PBTW exam. In addition, they used more sophisticated vocabulary and a wider variety of nouns in the PBTW exam. The students' essays did not vary in terms of syntactic complexity or grammatical accuracy, however. The scores that students received in the two exams correlated only moderately (.391), which suggests that the two exams measure different constructs, with the PBTW exam measuring reading, listening, and source integration, among other skills, in addition to writing. Both the learners and the raters mentioned that the learners had difficulties integrating sources, which could be a result of the negative washback of TW exams in their classes. The inter-rater reliability coefficient for all of the exams was high, but the inter and intra-reliability coefficients for the TW exam were higher. The results of this study, combined with the results of other similar studies and the skills that L2 learners need to succeed in college-level classes suggest that the PBTW exam may be a better tool to evaluate the construct of academic writing.

I dedicate this work to my husband, Eric Timothy David, and to my son, Liam Felipe David.

ACKNOWLEDGEMENTS

First and foremost, I thank all of the members of my committee, Dr. Charlene Polio, Dr. Peter De Costa, and Dr. Susan Gass, who have provided wonderful help throughout the data collection, analysis, and writing processes. I especially thank my chair and advisor, Dr. Paula Winke, who has guided me throughout my PhD program and given me incredible help and support. I especially thank *Language Learning* for awarding me with the Language Learning Dissertation Grant.

I also thank all of my family members, who have always supported me and believed in me, especially my parents, Edgar Harckbart and Solange Correa Harckbart, my brothers Tadeu Harckbart and Gustavo Harckbart, and my in-laws, Joan Elizabeth David and Patrick Timothy David. I am grateful to the many wonderful colleagues that I have met during my PhD program and the feedback that they have given me about this project and other projects. I have met wonderful instructors at the English Language Center and thank them for their wisdom and kindness, especially Carol Arnold and David Krise. I am so grateful to all of the students that I have had as a teaching assistant at Michigan State University. Teaching is my passion and I am lucky to have had the amazing students that I had. I am grateful to my friends in Brazil, in the United States, and elsewhere, who have always supported me. I especially thank Dr. Scott Sterling and his wife, Kara Sterling for their support and for the fun we have had together.

Thank you to my husband and best friend, Eric Timothy David, for your patience, understanding, and everlasting support. Thank you to my son, Liam Felipe David. Your smiles and laughter have given me much strength and happiness. I love you both.

TABLE OF CONTENTS

LIST OF TABLES

INTRODUCTION

Many English as a second language (ESL) writing programs all over the world use timed-writing exams for placement and achievement purposes. Placement exams are used to make decisions about which class or classes students should be placed in, and achievement exams assess whether students have mastered the goals and objectives of a course (Carr, 2011). Most timed-writing exams require students to write about a general topic for a specified amount of time, such as 30, 45, or 60 minutes. The essays are then scored by trained raters, and the test administrators use the scores to make decisions regarding whether students have to take writing classes, what should be taught in the classroom, or whether students are ready to move on to the next proficiency level. Timed-writing exams are extremely cost-effective, easy to design, administer, and score. However, timed-writing exams also have some disadvantages when administered for achievement purposes. Below I review five of the problematic areas.

First, if the purpose of an ESL writing program is to prepare students to succeed in an academic environment, impromptu timed-writing exams may not be the best tool to evaluate students' writing abilities. Many studies have found that the majority of academic writing tasks that students have to do in their regular university courses do not include answering bare prompts (Cooper & Bikowski, 2007; Hale et al., Horowitz, 1986; Yigitoglu, 2008). In a review of 54 writing assignments sheets from 29 different university-level courses, Horowitz (1986) found that the majority of assignments that students have to do in their academic courses involve the incorporation of sources. The most common assignments were: summary/critique, annotated bibliography, report on an experience, connecting theory and data, case study, synthesis of various sources, and research project. More recently, Cooper and Bikowski (2007), after

examining 200 university course syllabi, found that 38% of the assignments that students had to do were research papers, and 20% were book reviews. Impromptu timed-writing exams usually require students to write about their personal experiences or general topics, without allowing them to incorporate sources.

Hale et al. (1996) also investigated writing tasks that students have to perform at the university level. They analyzed writing tasks from 162 undergraduate and graduate courses in seven American universities and one Canadian one. Most of the in-class writing assignments in the undergraduate courses required students to write short essays of no longer than half a page. The most common out-of-class writing assignment was the research paper, multiple pages in length. Extending on Horowitz (1986) and Hale et al.'s (1996) studies, Yigitoglu (2008), in her master's thesis, collected 350 syllabi and handouts from different academic courses at a large Midwestern university to analyze the types of writing tasks that the professors required their students to do. The syllabi and handouts came from courses in the Social Sciences, Sciences, and Humanities. She found that the most common assignments were research and reaction papers. Forty-three percent of the writing prompts were text-based prompts, meaning that the students had to use sources in their writing, while only 29% were bare prompts, with no use of sources. Impromptu timed-writing exams usually require students to write about their personal experiences or general topics and do not provide them with readings to allow them to incorporate sources in their writing. This practice seems to go against a very robust finding that most university professors require students to incorporate sources in their writing.

Second, as Weigle (2002) noted, most writing assignments that students have to complete in regular academic classes are untimed and completed outside of the classroom. She further explained that timed-writing exams are not authentic academic writing tasks because, when

students write a paper for a course, the person who reads the essay is the professor, not a trained rater. The professor is usually more concerned with the paper's content, not grammatical accuracy, as a rater scoring timed-writing essays in an ESL course might be. Impromptu timed-writing exams also lack content validity when it comes to what many ESL academic writing teachers do in their classes. Many ESL programs that aim at preparing students for university academic writing teach writing as a process that includes reading, discussing, planning, peer review, revisions, and so on. A good example of this practice comes from books commonly used by ESL programs to teach academic writing. *Sourcework: Writing from Sources* (Dollahite & Haun, 2011), for example, has students engaging in research, summary writing, source integration, planning, and other skills that are not valued in impromptu timed-writing exams. Assessing learners using impromptu timed-writing exams might not provide program administrators with accurate information about what students learned during the course of a semester because these types of exams do not engage in the skills that students learn in their academic writing classes.

The third problem with impromptu timed-writing exams is the issue of topic familiarity. Research seems to suggest that when students write about a familiar topic they score higher (He & Shi, 2012; Tedick, 1990; Winfield-Barnes & Felfeli, 1982). When designing an impromptu timed-writing exam, it is really difficult to find a topic with which all students in a program, who come from many different cultural backgrounds, will be familiar. If students have to write about an unfamiliar topic, their scores could suffer.

Fourth, studies have also found that students perform better at writing tests if they are given time to plan what they will write (Ellis & Yuan, 2004; Kellogg, 1988; Worden, 2009). Most timed-writing exams require students to perform their best in a very limited amount of

time, with little time for planning. In an attempt to solve the problems that arise with impromptu timed-writing exams, more standardized ESL tests and programs are using integrated writing tasks (tasks that integrate reading and/or listening with writing) instead of, or in addition to, impromptu timed-writing exams. The TOEFL iBT, for example, in addition to an independent writing task, now includes a task for which students have to read a passage and watch a lecture to respond to a prompt (see www.toefl.org for more information). The ESL writing placement exam at the University of Illinois at Urbana-Champaign includes a mini-lecture, a reading passage, group discussions, and peer review (Cho, 2001). Research suggests that students score higher in integrated writing tasks than they do in independent writing tasks (Cumming et al., 2005; David, under review; Plakans, 2008). Integrated writing tasks provide students with the background information they need to write about the topic, even if they were initially unfamiliar with the topic.

The fifth disadvantage with impromptu timed-writing exams is that oftentimes students memorize phrases and even complete essays ahead of time, memorize them in entirety or in part, and use what they have memorized in writing exams. In a study examining students' perceptions of TOEFL's Test of Written English (TWE) (see www.toelf.org), many participants admitted to memorizing entire essays to prepare for the test (He & Shi, 2008). Because the prompts designed for impromptu timed-writing exams have to be about more general topics, test-takers have a good chance of memorizing an essay for a prompt they might actually encounter in a future test. Weigle (2013) argued that one advantage of integrated writing tasks is that they "counter test method or practice effects associated with conventional item types" (p. 2). Integrated writing tasks allow for a wider range of topics because the necessary background information that test-takers need to perform the task will be given through readings and/or videos. Moreover, students

are required to use the sources they read in their essays. These factors could decrease the change of students "regurgitating" essays in writing exams. To my knowledge, not many studies have examined students' perceptions of ESL writing exams. As stakeholders in exams for progress and achievement purposes, teachers and test designers should also take into consideration students' opinions of such tests. The students are the ones who spend hours preparing for the exams, taking the exams, and baring the consequences of their performance in the exams. Moreover, impromptu timed-writing exams do not mirror what students do in academic writing classes or other regular university classes, as described above. If the way that we assess learners in a program does not mirror what teachers actually do in the class, students may not feel invested in the exam and may not feel motivated to take the exam or perform well. In a prior study, I investigated students' perceptions of bare-prompt versus process-based writing tasks (David, under review). I found that students preferred process-based writing tasks over independent writing tasks. That study spurred me to delve into this topic more deeply. I started investigating the literature on L2 writing assessments more, and I found the five major problem areas in L2 writing assessment that I described above.

The general questions I have as an applied linguistics researcher and as an L2-writing teacher are these: Do students perform differently in writing exams when they are given (a) background readings and videos, (b) the opportunity to discuss their ideas in groups, and (c) time to plan their writing? What are students' perceptions of writing tests? What do raters think of writing tests? Before launching into the study at hand, I first review studies from researchers that have investigated these questions over the last 20 years. I review how complexity, planning, topic familiarity, writing-task integration, and test takers' perceptions have been researched from different angles. I also review how researchers have measured (quantitatively) writing success.

This is important because to compare essays that are written under varying conditions, researchers need an objective way to measure the essays using the same proficiency-oriented scale. Student perceptions and their thoughts, I have found, is not enough information to adequately compare performances. L2-writing researchers need both qualitative (interview response data) and quantitative measures (test scores) to understand the full scope of differences across exam formats. Thus, in this study, I use a sequential mixed-methods design (Creswell & Plano Clark, 2011) to compare the scores students receive in an impromptu timed-writing (TW) exam with the scores they receive in an integrated writing exam, which I call process-based timed-writing (PBTW) exam, and to investigate students' and raters' perceptions of both exams.

CHAPTER 1: REVIEW OF THE LITERATURE

1.1 Writing task complexity

L2-writing test designers have many options. They can allow for planning before the test takers write their essays; the writing topic can be familiar or unfamiliar; the writing task itself can be integrated (involve reading and/or listening); or not (bare prompt). In essence, test developers can manipulate the complexity of the writing task. Robinson (2001) explained that task complexity is "the result of the attentional, memory, reasoning, and other information processing demands imposed by the structure of the task on the language learner" (p. 29). He added that when a task is simpler, the learner will make less language errors, whereas when a task is more complex, the learner will be prone to making more language errors because the processing demands of the task are higher. According to Robinson (2001), there are different factors that affect the complexity of a task. If the learner is required to simply provide information as opposed to using reason, for example, the task is less demanding. At the same time, allowing planning time also makes the task simpler. However, if the learner is required to perform another task in addition to to the primary task, the task becomes more demanding. For instance, if the learner is required to read a text before writing, the task is more demanding, because the learner has to allocate resources to both tasks, not just one (Robinson, 2001). Background knowledge is another factor that can affect task complexity. If the learner has background knowledge for the task that he or she is completing, then the task is simpler. It is important to acknowledge that the issue of the effects of task complexity on L2 learners' oral and written performance is a controversial one, however. A meta-analysis conducted by Jackson and Suethanapornkul (2013) suggested that task complexity has only a slight influence on learners'

performance. Thus, task complexity's influence on performance may depend in large part on the context, including the performance-scoring system. Below I describe the research that has been done to investigate the effects of some of the factors of task complexity, such as planning, topic familiarity, and integrated tasks.

1.1.1 The effects of planning on L2 writing

Studies on the effects of planning on writing have mixed findings. While some researchers have found that students write higher quality essays when they are given time to plan (Ellis & Yuan, 2004; Kellogg, 1988; Worden, 2009), others failed to find this correlation (Johnson, Mercado & Acevedo, 2012; Ong & Zhang, 2013; Shi, 1998). Ellis and Yuan (2004) divided 42 L2 learners into three groups: pre-task planning, on-line planning, and no planning. In the no planning condition, the participants had to write an essay with a minimum of 200 words in 17 minutes. The participants in the pre-task planning group were given 10 minutes to plan their writing in addition to the 17 minutes to write their essay. In the on-line planning condition, the participants were not given a time limit or word limit to write their essay. The authors found that the pre-task planning group wrote longer essays with more syntactic variety, whereas the on-line planning group wrote more accurate essays.

The other study that showed a positive relationship between planning and the quality of writing was that of Kellogg (1988). Kellogg assigned 18 college students to two groups, which Kellogg further divided into two other groups: outline vs. no outline and rough draft vs. polished draft. The outline group was told to spend 5 to 10 minutes writing an outline for the writing task, whereas the no outline group was told to begin writing right away. Similarly, the rough draft group was told to write a rough draft of the essay, mainly to put their thoughts down on paper,

and then they were asked to revise their essay by adding and changing the content of their writing. The polished draft group was simply told to write their essay. Kellogg found that the outline group wrote faster than the no outline group. The outline group also wrote longer essays and received higher scores for their essays when compared to the no outline group. There were no significant differences in the efficiency with which the participants wrote their essays or in the quality of their writing for the rough draft and polished draft groups.

In another study investigating the effects of planning on learners' performance, Worden (2009) examined 890 essays written by L2 learners. Six hundred and forty students did some type of pre-writing activity and 747 students revised their essays in some way. The researcher found that the participants who were engaged in "high levels of pre-writing" received higher scores (p. 162). Participants who made global revisions to their essays, on the other hand, received lower scores.

Three studies, however, found no such correlation between planning and learners' writing performance. Shi (1998) had the participants engage in pre-writing discussions before they wrote their essays, but found no significant differences in the overall quality of the essays whose participants participated in such discussions and those who did not. Ong and Zhang (2013) found a positive relationship between planning time and fluency and lexical complexity. The 108 participants were given 10 minutes to plan and 20 minutes to write (pre-task condition); 20 minutes to plan and 10 minutes to write (extended pre-task condition); or 30 minutes to write continuously without any type of planning (free-writing condition). The authors found that the participants in the free-writing condition wrote significantly longer essays and scored significantly higher in lexical complexity than the other two groups, indicating that planning time does not affect fluency and lexical complexity.

In a large-scale study to investigate the effects of planning on L2 writing, Johnson, Mercado, and Acevedo (2012) examined the essays of 968 ESL students who were divided into five groups: a control group, an idea generation group, an organization group, a goal setting group, and a goal setting plus organization group. The control group did a vocabulary activity before they wrote. The idea generation group was given 10 minutes to list as many ideas related to the topic of the essay as possible. The organization group was given an outline worksheet before they wrote. The goal setting and the goal setting plus organization groups listed the rhetorical goals for the essay, but in addition to that, the latter also completed an outline worksheet before writing their essay. The results showed that planning had no effects on lexical complexity or grammatical complexity, but planning did result in longer essays, although the effect size was considerably small. One problem with this study was the fact that the researchers did not measure the level of the students' proficiency. Instead, they used the students' level in the ESL course as a measure of proficiency. Students were from four different Advanced levels. As Johnson et al. explained, the participants' level of proficiency could have influenced their ability to free cognitive resources for successful planning. They suggested the following:

> [There may be] a threshold of proficiency in the target language in order for working memory resources to be freed from the demands of the translation process of writing to such an extent that pre-task planning may have any measurable impact on features of L2 writers' texts. (p. 272).

Although some of the studies mentioned above did not find any positive effects on planning on L2 learners' writing, others have found that planning results in longer essays, higher scores, and sometimes more syntactic variety (Ellis & Yuan, 2004; Kellogg, 1988; Worden, 2009). These results suggest that learners could benefit from being given time to plan what they

will write. These findings also align with Robinson's (2001) idea that planning reduces the complexity of a task.

1.1.2 Topic familiarity

Most studies that have investigated the effect of topic familiarity on writing have shown that L2 learners write more fluent essays and score higher when they write about topics with which they are familiar (He & Shi, 2012; Tedick, 1990; Winfield-Barnes & Felfeli, 1982). He and Shi (2012) asked 50 language learners to write two essays, one in response to a prompt about a general topic that required them to use personal experiences, and another in response to a specific topic. The prompt that they used for the general topic was "If you plan to attend college or a university, what factors will influence your choice of what you study? Provide reasons." The specific topic prompt was "Explain why you do OR do not take an interest in federal politics. Be specific." The participants performed significantly better on the general topic than on the specific topic. The authors explained that the participants wrote shorter essays with weaker cohesion and coherence and more grammatical errors on the specific topic. In addition, the participants' essays lacked idea development, and the participants did not explicitly state their position on the issue.

Seeking to examine the effects of topic knowledge on writing performance, Tedick (1990) collected two writing samples from 105 ESL students. For one essay, the participants wrote about a general topic, and for the other essay, the participants wrote about a topic related to their fields of study. When they wrote about their field of study, the participants received higher holistic scores and made fewer grammatical mistakes. Tedick concluded that topic familiarity is related to better performance in writing. The same conclusion was reached in another study (Winfield-Barnes & Felfeli, 1982).

In Winfield-Barnes and Felfeli's study, ten of the twenty participants were from a Spanish speaking country and the other ten were from other non-Spanish speaking countries. All of the participants were asked to read two paragraphs about the book *Don Quixote* and a Japanese play named *Noh*. They were then asked to write two compositions about the two paragraphs. The authors noticed that the Spanish-speaking participants wrote longer and more accurate essays when they wrote about the *Don Quixote* paragraph. They concluded that "easing the dual cognitive processing load by having students deal with culturally familiar material increases fluency" (p. 376).

The findings of the studies described above seem to indicate that when L2 learners are familiar with a topic, they write higher quality essays and score higher (He & Shi, 2012; Tedick, 1990; Winfield-Barnes & Felfeli, 1982). Again these findings align with Robinson's (2001) theory. When learners write about familiar topics, they have more attentional resources to allot to other elements of writing, such as cohesion and grammar.

1.1.3 Integrated tasks

Many researchers have investigated tasks that integrate reading and/or listening with writing. However, the results are conflicting. Gebril (2010), for example, administered two independent writing tasks and two reading-to-write tasks to 115 English as a foreign language (EFL) learners in Egypt. He found that the participants' scores in the two sets of tasks correlated highly and suggested that the two tests measured similar constructs. However, the findings of other studies suggest the opposite: Integrated writing tasks and independent writing tasks measure two different constructs (Cumming et al., 2005; David, under review).

Cumming et al. (2005) examined test takers' discourse in independent and integrated writing tasks written for the TOEFL (www.toefl.org) and found that the discourse differed significantly in terms of text length, lexical complexity, syntactic complexity, and discourse orientation. The test takers wrote shorter essays, used a wider variety of words, and wrote longer clauses and more clauses in the integrated writing task. In addition, they tended to use less personal knowledge and more source information. However, the authors noticed that the test takers' essays were more argumentative in nature in the independent writing tasks. The essays did not differ significantly in terms of grammatical accuracy. However, when the researchers analyzed the more advanced learners, they found somewhat different results. The more advanced learners wrote longer essays, used a wider range of words, wrote longer clauses and more clauses, were more grammatically accurate, and provided better arguments in the integrated writing tasks. It seems, then, that proficiency plays a part in the two types of tests; more advanced learners seem to take more advantage of integrated writing tasks, or it could be that they are able to better demonstrate their advanced proficiency through integrated-task work.

Plakans (2008) investigated the differences between ten test takers' processes in independent and reading-to-write writing tasks. Nine of the 10 test takers believed that they performed better in the integrated writing task and they were indeed right in doing so. The writing process that they used, however, differed. The participants planned more in the independent writing task. Furthermore, more experienced writers seemed to engage more with the text in the reading-to-write task.

A study of much relevance to this is one in which I compared students' performance and perceptions of a process-based writing (PBTW) exam and an impromptu timed-writing (TW) exam (David, forthcoming). In my study, 40 students from Academic writing classes in a large

Midwestern university took the two exams within the same week. I counter-balanced the order of the administration of the exams. The PBTW exam required students to read two texts and watch two videos about the same topic. After that, the participants were presented with the essay prompt and given 10 minutes to discuss their ideas in groups. In addition, the participants had 10 minutes to plan and 45 minutes to write their essays. For the impromptu TW exam, the participants had to choose between two prompts and write their essays in 45 minutes. Two raters scored each essay using a holistic rubric. The results revealed that students scored higher in the PBTW exam. Although a *t* test did not reveal a significant difference between the scores in the two exams, the results approached significance ($p = .06$). Furthermore, the scores the students received in the two exams correlated only moderately ($r = .417$), indicating that the exams probably measure different constructs. The participants also wrote significantly longer essays in the PBTW exam and used more verb phrases per T-unit than in the impromptu TW exam. There were some limitations to this study, however. The participants had a choice of topic in the TW exam, and the two topics they could choose from in the exam might not have been comparable. One topic required students to write using personal experiences, while the other was a more academically oriented topic about the use of natural resources in the world. The rubric used to rate the PBTW exam was the same as the one used to rate the TW exam, with one exception. I added a column to the rubric to account for the use of sources in the participants' writing. Adding another component to the rubric was a problem because some participants were penalized for not using the required sources. In addition, the raters only used a holistic rubric to rate the essays. An analytic rubric might provide more fine-grained and specific information about the participants' scores.

Plakans (2008), Cumming et al. (2005), and David (forthcoming) found very similar results in their studies: The participants scored higher and wrote more syntactically complex essays when they were given reading and listening passages related to the topic of the essay. Furthermore, the results of Cumming et al.'s study suggested that integrated tasks may allow more advanced learners to write more accurate essays.

1.1.4 Test takers' perceptions of writing tests

Test designers need to consider more than just the dimensions of the writing test when they choose how to formulate and construct a writing test. They must also consider the population of test takers. What will the test takers think of the test? This is relevant because "they are important stakeholders (perhaps the most important) but their views are among the most difficult to make sense of and to use" (Rea-Dickins, 1997, p. 306). As explained by McNamara and Roever (2006), the International Language Testing Association (ILTA) detailed in its Code of Ethics (ILTA, 2000) that a fundamental concern of any test developer should be the protection of test takers, who are in the testing process's least powerful position. At the same time, ILTA recognized within its Code of Ethics that test developers are often pulled between their need to assess in efficient and expedient manners and their need to be contentious of the test takers and their value systems. As explained above, the test takers are the ones preparing for the exam, taking the exam, and baring the consequences of the exam. If, for some reason, students are not invested in an exam, they may perform poorly due to lack of motivation (Lee & Coniam, 2013). However, to my knowledge, there are a few studies that have examined test takers' perceptions of writing tests (David, under review; He & Shi, 2008; Lee, 2006; Powers & Fowles, 1999). One of these studies sought to examine test takers' perceptions of two standardized

writing tests: TOEFL'S Test of Written English (TWE) (www.toefl.org) and the English Language Proficiency Index (LPI) (see www.celpiptest.ca). The TWE includes an integrated writing task and an independent writing task. In addition to the TWE, many international students in Canada are required to take the LPI, a Canadian English proficiency test.

He and Shi (2008) interviewed 16 Chinese students attending a Canadian university about how they prepared for the two tests and about their perceptions about both tests. Of the 16 participants, 13 failed or had difficulties passing the LPI. With the exception of two participants, all participants stated that they prepared for the TWE by memorizing sentences, and some even admitted to memorizing entire essays. To prepare for the LPI, the participants took a 3-month preparatory course at the university. The course focused on writing skills, and not memorization, and most students felt that the course did not help them pass the LPI. The authors attribute the participants' negative perceptions about the preparatory course to how they are used to preparing for writing tests in their home countries: through memorization. Most of the participants believed that the LPI was more difficult than the TWE because they had to write about unfamiliar topics. Furthermore, they also felt that they needed to know more about the Canadian culture to write about the topics in the LPI. They thought that the TWE, on the other hand, provided them with more general topics. Finally, He and Shi reported that the participants felt more pressure to write grammatically accurate essays in the LPI than in the TWE. In the TWE, students had memorized sentences and essays, which helped them to write more accurate essays.

In another study regarding test-takers perceptions of writing exams, Powers and Fowles (1999) investigated students' opinions of possible prompts for the Graduate Record Examination (GRE, see www.ets.org/gre). First, the researchers asked the participants to rate six prompts on a 7-point scale, in which 7 was extremely good, and 1 was extremely poor. They were also asked

with which prompt they could write the best and the worst essay and to explain why. After rating the essays, the participants had to write two essays using two of the six prompts they rated. After they wrote the essays, the participants were asked to rate the six prompts once again. The researchers found that the participants, as individuals, were not very consistent with their ratings of the six prompts. In other words, they rated the prompts differently. However, they seemed to be consistent when considered as a group. On the one hand, the participants thought that a prompt was good when: 1) they could connect with the topic; 2) the prompt was clearly explained; and 3) they thought the topic was interesting. On the other hand, the participants rated prompts as difficult when: 1) the topic was not interesting; 2) they lacked topic familiarity; 3) the prompt was not clearly explained; 4) the participants had negative feelings about the topic; and 5) they thought the topic was not pertinent. However, some participants rated a prompt as easy, while others rated the same prompt as difficult, indicating that participants have different opinions about the prompts. Finally, Powers and Fowles compared the essays that the participants wrote for prompts they rated as difficult to those they rated as easy. The authors found that the participants tended to do better when they wrote essays for the prompts that they rated as easier. The results of this study seem to suggest that students' perceptions about prompts are related to their performances.

Lee (2006) investigated a daylong ESL writing placement exam at the University of Illinois at Urbana-Champaign. The ESL learners watched a mini-lecture, read an article, participated in group discussions, had time to plan, write, and then peer review their essays. The author included a short survey about the students' perceptions of the placement exam at the end of the day. The students believed that the exam elicited their "true" writing abilities. While some

participants complained that the exam was too long, others said the opposite. The participants also made positive comments about the group discussions and peer feedback.

In the study that I conducted comparing two types of writing exams (David, under review) I also investigated the test takers' perceptions of the two exams. After the participants took the two exams, they answered a short questionnaire that contained both multiple-choice and short-answer questions about the two exams. Seventy percent of the participants preferred taking the PBTW exam and 62% thought that it was easier than the impromptu TW exams. Six participants also said that they derived ideas from the group discussions in which they participated. Almost half of the participants (n = 18) said that they gleaned ideas from the articles they read and the videos they watched and used these ideas in their writing. Some participants (n = 8) mentioned that they liked the fact that they had time to plan their essays before they began writing.

My study (David, forthcoming) and Lee's (2006) study indicated that learners value the process of writing through which process-based writing exams allow them to go. The participants in both studies reported that they liked having time to plan their essays and discuss their ideas in groups. Furthermore, the participants in my study reported that they learned ideas from the topic by reading and watching videos and used those ideas in their writing. Powers and Fowles (1999) found that the students' preferences of prompts may be related to their performance, which could indicate that indeed test takers' perceptions of exams may have an effect on their performance. If students do not like the exam that they are taking, they may not feel motivated to take it or do well (Lee & Coniam, 2013).

1.2 Measuring the components of writing

Above I reviewed writing task complexity, which is important for test development, and test takers' perceptions of writing exams, another crucial component of test design. In addition to task complexity and test takers' perceptions of writing exams, it is important for test designers to determine which components of writing to evaluate them and how to measure each of these components. The most commonly used holistic and analytic rubrics contain categories like content, organization, grammar, vocabulary, and so on. Jacobs, Zinkgraf, Wormuth, Hartfiel, and Hughey's (1981) rubric, for example, includes content, organization, vocabulary, language use, and mechanics. Weir's (1990) rubric divides mechanics into spelling and punctuation and also includes cohesion. Task achievement, coherence and cohesion, lexical resource, and grammar are the categories one can find in IELT's rubric for one of the test's writing tasks. The researchers mentioned above who have investigated the effects of planning, task complexity, and integrated tasks on learners' writing have compared test takers' writing using the following dimensions: lexical and syntactic complexity, fluency, and grammatical accuracy (Cumming et al., 2005; Ellis & Yuan, 2004; Johnson et al., 2012; Shi, 1998). Other researchers have also analyzed test takers' writing using the same measures. Below I review how these researchers have analyzed grammatical accuracy, lexical and syntactic complexity, and fluency.

1.2.1 Measures of grammatical accuracy

Most of the researchers who have analyzed grammatical accuracy in L2 writing used measures such as error-free T-units or error-free clauses. T-units are independent clauses and their accompanying dependent clauses (Hunt, 1965). Researchers decide, on their own, what "error-free" means when counting error-free T-units, but many use Polio's (1997) guidelines to

do so. For Polio's guidelines on how to count error-free T-units, see Appendix F. Amstrong (2010), Kuiken, Mos, and Vedder (2005), and Wigglesworth and Storch (2009) measured grammatical accuracy by calculating the percentage of error-free T-units per total T-units. Amstrong (2010) investigated the effects of giving grades for compositions written by intermediate L2 Spanish learners on fluency, accuracy, and complexity. The author defined accuracy as "the errors in grammar and vocabulary" (p. 693) and she compared students' performance on essays that were graded, essays that were ungraded, and written work done for an online discussion board. Amstrong measured accuracy by analyzing the number of error-free T-units, error-free T-units per T-unit, and errors per T-unit. She only found a significant difference for errors per T-unit between the graded and ungraded essays and for error-free T-units between the ungraded essays and the online discussion board. There were more error-free T-units in the online discussion board. Kuiken et al. (2005) also used error-free T-units per total T-units as a measure of grammatical accuracy, as well as dividing the degree of errors in three categories: minor errors, more serious errors, and errors that make the text incomprehensible. The authors investigated the effect of task complexity on syntactic complexity, lexical variation, and accuracy. The participants were 62 Dutch learners of Italian who performed two writing tasks, one of which was more cognitively complex than the other. The results of the study showed that the learners made significantly more errors in the more complex task. Another study that measured grammatical accuracy was that of Wiggleswroth and Storch (2009). The authors wanted to investigate whether advanced L2 learners' writing differed in terms of grammatical accuracy, fluency, and complexity when students worked in pairs or individually. In addition to error-free T-units per total T-units, they also calculated the percentage of error-free clauses. The results of the study revealed that the 144 participants produced significantly less error-free T-

units when they wrote the essays in pairs. Ellis and Yuan (2004) calculated the percentage of error-free clauses per total clauses in the study mentioned above. The authors investigated the effects of planning in L2 writing on fluency, complexity, and accuracy. They had two measures of accuracy: error-free clauses per total clauses and correct verb forms, that is, the percentage of verbs used accurately. For the first measure of accuracy, the researchers counted as errors syntactic, morphologic, and lexical choice mistakes. As mentioned above, the learners in the on-line planning condition wrote essays with significantly less grammar errors than the learners in the no planning and pre-task planning conditions. Of particular interest to the present study is the work by Polio and Shea (2014). Polio and Shea investigated different measures of linguistic accuracy. The authors analyzed the following measures of linguistic accuracy found in 35 different studies in an attempt to investigate their reliability: holistic scores of language use and vocabulary, error-free T-units per total T-units, error-free clauses per total clauses, weighted error-free T-units per total T-units, number of errors per words, and number of verb phrase, preposition, article, and lexical errors per words. Polio and Shea used these measures of grammatical accuracy to analyze essays in the MSU data set. This data set includes essays written over the course of a semester by ESL students enrolled at the English Language Center at Michigan State University. The inter-rater reliability for error-free T-units per total T-units and error-free clauses per total clauses was .88; .85 for language holistic scores and .90 for vocabulary holistic scores; .84 for weighted error-free T-units; and .89 errors per total words. The other measures of linguistic accuracy all had inter-rater reliability lower than .80.

Error-free T-units and error-free clauses seem to be among the most commonly used measures to investigate accuracy in L2 writing and researchers have found that these are both reliable measures to do so (Polio and Shea, 2014).

1.2.2 Measures of lexical complexity

Many researchers have analyzed L2 learners' essays in terms of lexical complexity, and there are quite a few lexical complexity measures from which to choose. Fritz and Ruegg (2013) investigated whether raters are sensitive to lexical accuracy, lexical sophistication, and lexical range when rating timed writing essays. Eight hundred and ninety-five EFL learners wrote an essay responding to a single prompt in 30 minutes. The researchers manipulated these essays in terms of lexical accuracy, sophistication, and range. The authors included two types of errors in the essays: errors of word choice, that is, using the wrong word in a certain context, and errors of part of speech, that is, using the wrong part of speech of a word in a certain context. The manipulated essays contained three levels of accuracy: low accuracy, when the essay had errors in 32 content words; medium accuracy, when there were 16 errors in the 32 content words; and high accuracy, when all of the 32 content words were correct. The authors used RANGE, developed by Nation (2005), to manipulate lexical sophistication. RANGE compares essays to three different word lists: the most common 1,000 words (1,000 word level), the second most common 1,000 words (2,000 word level), and the third most common 1,000 words (3,000 word level). Fritz and Ruegg created essays with three levels of lexical sophistication: low lexical sophistication (32 content words from the 1,000 word level), medium lexical sophistication (32 content words from the 2,000 word level), and high lexical sophistication (32 content words from the 3,000 level). The researchers created three categories when they manipulated lexical range: high lexical range, when the essay contained 32 different content words; medium range, when the essay contained 25 different content words; and low range, when the essay had 18 different content words. Twenty-seven raters scored the manipulated essays using a four-point analytic

rubric. The results of the study revealed that raters seem to be sensitive to lexical accuracy, but not to lexical range or lexical sophistication.

Another study that measured L2 learners' lexical complexity was by Kormos (2011). Kormos investigated how task complexity influenced L2 learners' lexical diversity and lexical complexity. In addition, the author also compared L1 and L2 writers' essays in terms of lexical complexity. The participants performed two writing tasks: a picture description task and a picture narration task. The main difference between the two tasks is that the latter is said to be more complex because "the participants not only had to rely on their language skills, but they also had to use their imagination and find a way to relate the pictures to one another and invent a story around them" (Kormos, 2011, p. 153). Kormos used Malvern and Richards' (1997) *D*-formula and Nation's (2005) RANGE program to measure lexical range, in the same manner Fritz and Ruegg (2013) used the program. One more measure of lexical complexity was used in this study: the concreteness of the words, that is, how concrete or abstract words are in a text. The results of the study show that task complexity did not influence lexical complexity at all. One reason for that might be that the two writing tasks were in the same genre, narratives. Kormos did, however, find a significant difference in lexical variety and complexity between L1 and L2 writers. A study similar to Kormos' was that of Kuiken et al.'s (2005), mentioned above. In addition to analyzing how task complexity influenced accuracy in L2 writing, the authors also investigated lexical variation. They measured lexical variation as the number of word types divided by the total number of word tokens and word types per square root of two times the total number of word tokens. Similarly to Kormos' study, the results of Kuiken et al.'s study suggest that task complexity has no effect on lexical variation.

With the aim to analyze the relationship between lexical richness and raters' judgments of ESL learners' spoken language, Lu (2012) created the Lexical Complexity Analyzer. The Lexical Complexity Analyzer generates 25 different measures of lexical complexity that measure three different dimensions of lexical complexity: lexical density, lexical sophistication, and lexical variation. Lexical density, according to the author, "refers to the ratio of the number of lexical (as opposed to grammatical) words to the total number of words in a text" (Lu, 2012, p. 191). The author included the following categories in his definition of lexical words: nouns, adjectives, verbs, and adverbs. For lexical sophistication, Lu analyzed five different measures: lexical sophistication 1, following Linnarud (1986) and Hyltenstam (1988), who divided the total number of sophisticated lexical words by the total number of lexical words in a text; lexical sophistication 2, following Laufer (1994), who analyzed lexical sophistication in terms of the number of sophisticated words divided by the number of words in a text; verb sophistication 1, following Harley and King (1989), which is the number of sophisticated verbs divided by the number of total verbs in a text; verb sophistication 2, which is a modification of verb sophistication 1, following Chaudron and Parker (1990); and corrected verb sophistication 1, which is also an adaptation of verb sophistication 1, following Wolfe-Quintero, Inagaki and Kim (1998) (See Lu, 2012 for a more detailed explanation of these measures). Finally, Lu (2012) used 20 different measures to investigate lexical variation. These measures are the following: number of different words, first 50 words, expected random 50, expected sequence 50, type-token ratio, mean segmental TTR (50), corrected TTR, root TTR, bilogarithmic TTR, uber index, D measure, lexical word variation, verb variation 1, squared verb variation 1, corrected verb variation 1, verb variation 2, noun variation, adjective variation, adverb variation, and modifier variation. The participants in Lu's (2012) study performed three spoken tasks and trained raters scored their

performance on these tasks. The results revealed that lexical density and lexical sophistication did not correlate very strongly with the scores that the participants received. However, lexical variation strongly correlated with the participants' scores.

Two of the studies that investigated the effects of planning on writing mentioned above analyzed the effects of planning on lexical complexity (Johnson et al., 2012; Ong & Zhang, 2013). In Johnson et al.'s study, the researchers analyzed five different measures of lexical complexity: lexical diversity (using McCarthy's computer program and Coh-Metrix), the number of pronouns to noun phrases, the use of personal pronouns, how often content words appear in comparison to a corpus, and how often the four and five most frequent word families appear in comparison to a corpus (using Nation's (2005) RANGE). As mentioned above, Johnson et al. found no significant differences in terms of lexical complexity in the different planning conditions. Ong and Zhang (2013) used the following formula to calculate lexical complexity: WT^2/W, which means "word types squared divided by the total number of words" (p. 223). As mentioned previously, the results revealed that the participants who in the pre-task (10 minutes to plan and 20 minutes to write) and free-writing (write nonstop for 30 minutes) conditions wrote more lexically complex essays than the participants in the extended pre-task (20 minutes for planning and 10 minutes for writing) and control conditions (write for 30 minutes).

Some common measures of lexical complexity used by researchers to analyze L2 writing or speech are: lexical sophistication; lexical density; and lexical variety. Of particular interest to this study are the two programs developed by Nation (2005) and Lu (2012): Nation's RANGE analyzes lexical sophistication that compares essays to three different word lists; and Lu's Lexical Complexity Analyzer that examines 25 different measures of lexical complexity, all of which include lexical density and variety measures.

1.2.3 Measures of syntactic complexity

Many of the studies that investigated integrated tasks and the effects of planning on L2 writing included measures of syntactic complexity (Cumming et al., 2005; David, under review; Ellis & Yuan, 2004; Kuiken et al., 2005; Kuiken & Vedder, 2008). When comparing integrated writing tasks to independent writing tasks, Cumming et al. (2005) analyzed syntactic complexity in two ways: by counting the number of clauses per T-unit and the number of words per T-unit. The findings of the study revealed that the participants wrote significantly less words per T-unit in the listening to write tasks than in the reading to write and independent writing tasks. Ellis and Yuan (2004) measured syntactic complexity the same way Cumming et al. did: by counting the number of T-units per clauses. The findings of the analysis revealed that the participants who engaged in pre-task planning had significantly more syntactic variety than the no planning group. Another study that investigated the effects of task complexity on syntactic complexity was that of Kuiken et al.'s (2005). The authors used the following measures of syntactic complexity: number of clauses per T-unit and number of dependent clauses per clause. The results revealed no significant differences in syntactic complexity in the two tasks. In a later study, Kuiken and Vedder (2008) used similar measures of syntactic complexity to investigate the effects of cognitive task complexity on Italian and French as a Foreign Language writing. They calculated the number of clauses per T-unit and the number of dependent clauses per clauses. The authors found no significant differences in syntactic complexity. One study of particular interest to the present study is that of Ai and Lu's (2013). The authors investigated syntactic complexity in a large number of essays written by university-level native and nonnative speakers of English and they used the L2 Syntactic Complexity Analyzer, developed by Lu (2010) to do so. Ai and Lu

used ten out of the fourteen syntactic complexity measures generated by the L2 Syntactic Complexity Analyzer: mean length of clause, mean length of sentence, mean length of T-unit, dependent clauses per clause, dependent clauses per T-unit, coordinate phrases for clause, coordinate phrases per T-unit, T-units per sentence, complex nominals per clause, and complex nominals per T-unit. The authors found that the native speakers and nonnative speakers' compositions differed significantly in terms of syntactic complexity. I also used the L2 Syntactic Complexity Analyzer to compare L2 writers' essays in the PBTW exam and the impromptu TW exam. The results revealed that the participants wrote more verb phrases per T-unit in the PBTW exam, while they also wrote more coordinate phrases per clause in the TW exam.

Number of clauses per T-unit and number of words per clause or T-unit seem to be popular measures to investigate syntactic complexity in L2 writing. Lu's (2010) L2 Syntactic Complexity Analyzer includes these measures, as well as many others that are related to syntactic complexity, such as the number of dependent clauses per T-unit. Moreover, Lu's program has been found to be a reliable measure of syntactic complexity when compared to human raters (Lu, 2010).

1.2.4 Measures of fluency

Researchers analyzing fluency in L2 writing have measured the concept of fluency in different ways and there is much controversy on what is an accurate measure of fluency (Abdel Latif, 2012). Ellis and Yuan (2004), for instance, used two measures of fluency: syllables per minute and number of disfluencies. The authors defined the latter measure as "the total number of words a participant reformulated (i.e., crossed out or changed) divided by the total number of words produced" (p. 71). The results of Ellis and Yuan's study revealed that the participants

wrote more syllables per minute in the pre-task planning group than the no planning group. Johnson et al. (2012), on the other hand, investigated fluency by calculating the total number of words and average sentence length. The authors only found a significant different in the average sentence length. The learners in the control group wrote longer sentences than the learners in the pre-task planning condition. Similarly, Cumming et al. (2005) calculated the total number of words in participants' essays to measure fluency. This, however, was the only measure of fluency that the authors used. They found, however, that the participants wrote more in the independent writing tasks. The study that I conducted comparing PBTW exams and impromptu TW exams also included one measure of fluency (David, under review). In order to measure fluency, I divided the total number of words per essay by the total number of minutes that the participants were allowed to write. The results revealed that the participants wrote significantly longer essays in the PBTW exam when compared to the TW exam. This measure of fluency, however, was problematic because some participants finished writing before the 45 minutes were over and the number of words per minute was possibly not a true measure of fluency for those participants who finished writing earlier. Another one of the studies mentioned above that investigated fluency in L2 writing was that of Kellogg's (1988). The researcher measured fluency in terms of the total number of words, total time spent on task, and words per minute; a much more complete and accurate measure of fluency than the one used in my study (David, under review) study. The L2 learners in the outline condition wrote more words, spent more time on task, and wrote faster than the learners who did not write an outline for their essays. Ong and Zhang (2013) also investigated the effects of task complexity on fluency. The authors calculated fluency by analyzing words per minute for the total number of minutes on task (fluency II) and words per minute for the amount of time the learners wrote their essays (fluency I). The results

of the study revealed that the learners scored significantly higher for fluency II in the free-writing condition than the participants in the pre-task and extended pre-task conditions. However, there were no significant differences for the different planning conditions in terms of fluency I.

There are many problems with theses measures of fluency, as explained by Abdel Latif (2012). Abdel Latif stated that one of the problems with using any of the above measures of fluency is that writers often pause while writing and the pauses are not very consistent. In fact, Flower and Hayes (1981) said that more than half of the time that writers spend on task consists of pause time, not writing time. Total number of words per essay and words per minute do not take those pauses into consideration and therefore may not really reflect fluency accurately. Another problem that Abdel Latif reported with the most commonly used measures of fluency is that, when writers pause, they pause for different reasons. They may pause to plan, to monitor language, to retrieve information, and so on. The author explained that pausing can help or interfere with writing fluency.

CHAPTER 2: THE PRESENT STUDY

A need exists in the literature to develop a more thorough understanding of the effect of planning, topic familiarity, and integrated tasks on L2 writing, as well as test takers' and raters' opinions of writing exams by collecting both quantitative and qualitative data, because each type of data provides a different view of these issues. According to Creswell and Plano Clark (2011), a mixed methods study "focuses on collecting, analyzing, and mixing both quantitative and qualitative data in a single study or series of studies" (p. 5). They further explained that "its central premise is that the use of both quantitative and qualitative data approaches, in combination, provides a better understanding of research problems than either approach alone" (Creswell & Plano Clark, 2011, p. 5). The quantitative results from this study will provide statistical information about the test takers' performance in the two exams, while the qualitative results may provide explanations and insights about their performance, which will ultimately enhance and add depth to the findings of the study. The present study is a fixed mixed-methods study, rather than emergent and evolving, because the use of quantitative and qualitative methods for collecting data for the study were planned before data collection began (Creswell & Plano Clark, 2011). The data collection for this study was sequential (Creswell & Plano Clark, 2011), that is, data collection for the quantitative portion of the study was done first, by administering tests, followed by the collection of the qualitative data, which consisted of questionnaires and interviews. In addition, the results of both the quantitative and qualitative portions of the study were combined during the interpretation phase, after all of the data were collected. Creswell and Plano Clark (2011) explained that there are multiple ways to design and conduct mixed methods research: the convergent parallel design, the explanatory sequential design, the exploratory

sequential design, the embedded design, the transformative design, and the multiphase design. As I described, I followed the convergent parallel design, in which the researcher combines the results of the quantitative and qualitative data after the data for each portion of the study has been analyzed. The reason why this mixed-methods design was chosen is because, as Morse (1991) explains, the researcher can choose the convergent design when he or she wants "to obtain different but complementary data on the same topic" (p. 122). I wanted to gather information about the test takers' performance on the two exams, which can most commonly be done by using quantitative methods, but I also wanted to investigate the test takers' and the raters' opinions of the exams. The nature of qualitative research allows for a richer understanding of people's opinions. In addition, the convergent design allows the researcher to compare and contrast quantitative and qualitative results to validate the findings of each result and to have a better understanding of what they mean (Creswell & Plano Clark, 2011).

In an attempt to expand on and combine the findings of the studies mentioned in the literature review, the purpose of this mixed methods study is to investigate how test takers perform in two timed writing exams: an impromptu TW exam and a PBTW exam. Another goal of the study is to examine test takers' and raters' perceptions of the two exams. In the present mixed methods study I seek to investigate the following research questions:

1) Do test takers' scores in an impromptu TW exam correlate with their scores in a PBTW exam?
2) Do test takers' scores in an impromptu TW exam differ significantly from their scores in a PBTW exam?
3) How do students' writing across exams differ in terms of:
    a) Accuracy?

b) Lexical complexity?

c) Syntactic complexity?

d) Fluency?

4) What are the intra and inter-rater reliability coefficients for each exam? (Are they comparable?)
5) What are students' perceptions of the two different types of writing exams?
6) What are the raters' perceptions of the two different exams?

The quantitative data in this study are the test takers' scores on the exams, their accuracy, lexical and syntactic complexity, and fluency scores on the exams, as well as the intra and inter-rater reliability for the two exams. The qualitative data consist of the questionnaire and interviews. The qualitative data was collected and analyzed borrowing from the case study tradition. With case studies, "the focus is a small number of research participants" and "the individual's behaviors, performance, knowledge, and/or perspectives are then studied very closely and intensively, often over an extended period of time, to address timely questions about language acquisition, attrition, interaction, identity, or other current topics in applied linguistics" (Duff, 2012, p. 95). In the case of the present study, the focus was on 18 participants, but data collection did not occur over an extended period of time. Instead, the participants were interviewed only once after they took the two exams. Because one of the aims of the study was to investigate test takers' perceptions of the two exams that they took, it seemed unnecessary to conduct multiple interviews. Duff explained that researchers can learn much about individuals by examining a smaller number of them instead of a large number of participants. The individuals that I investigate are 18 out of the 81 participants and the two raters who participated in this study. These participants were either randomly selected by me, in the case of the two classes I taught,

selected by their teachers, or they volunteered for the interviews (in the case of the teachers who collected data for me). This particular investigation of the 18 test takers and the two raters is interpretive in nature. I seek to understand the test takers' and raters' perceptions of the two exams through interviews. The themes that emerged in the interviews with the test takers are described within the context of the responses of the post-writing questionnaire, which contained both multiple-choice and open-ended questions. Instead of describing the data case by case, I describe the themes that emerged across cases for both the test takers and the raters.

2.1 Method

2.1.1 Participants

The participants of this study were 81 ESL learners who were taking an academic writing course at the English Language Center at Michigan State University at the time of data collection, which occurred in the summer and fall semesters of 2014. There are two academic writing courses at the English Language Center (ELC): one is a 6-credit-hour course that focuses on grammar and writing (ESL 220); and the other is a 3-credit-hour course that focuses only on writing (ESL 221). Students are placed in these courses based on the score that they receive on a placement exam upon arrival on campus. Some students in ESL 220 and ESL 221 have taken other ESL courses at the ELC, while others are placed directly in these courses and have not taken other courses at the center. The participants in this study are from both ESL 220 and ESL 221. The reason why students from both courses participated is because both courses focus on academic writing, but ESL 221 does not focus on grammar instruction. Only eight of the participants were enrolled in ESL 221. I sent an email to the teachers who were teaching ESL

220 and ESL 221 during the summer and fall semesters in 2014 to ask them for help to collect data. Four teachers replied and agreed to allow me to collect data during class time. Two of the teachers agreed to do the data collection themselves because my schedule did not allow me to come into their classrooms to collect data. Two of the six classes in which the data were collected were taught during the summer semester by me. Some teachers decided to give extra credit to the students who agreed to participate in the study and others did not. Some teachers told their students that they would grade the students' essays as part of their coursework, and others did not use the essays for grading. The data were collected in five ESL 220 classes and one ESL 221 class. Table 1 summarizes the information about the participants. Fifty males and 31 females participated in the study. The average length of residence (LoR) in the United States of the participants was 10 months and the average number of years they had had formal English instruction (FI) was five and a half years. Approximately 38% (n=31) of the participants were from Brazil, 36% (n=30) were from China, 1% (n=7) were from Saudi Arabia, and the remaining number of participants were from Angola (n=5), South Korea (n=3), the United Arab Emirates (n=2), Kuwait (n=1), Lybia (n=1), and Japan (n=1). The most common majors among the participants were Business (n=7), Finance (n=6), Electrical Engineering (n=6), Accounting (n=5), Medicine (n=5), Agricultural Sciences (n=5), Civil Engineering (n=4), and Forestry (n=4). Many of the participants were already taking classes in their majors alongside their ESL classes, but some were only taking ESL classes at the time of data collection. Only two of the participants had undecided majors. The great majority of the students were undergraduate students. Five were graduate students. The median age of the participants was 21 years of age

Table 1
Participants

Country	**N**	**Mean age**	**Mean LoR**	**Mean FI**
Brazil	31	22	4 months	4 years
China	30	20	15 months	7 years
Saudi Arabia	7	25	13 months	5 years
Angola	5	20	9 months	1 year
South Korea	3	24	25 months	9 years
United Arab Emirates	2	18.5	9 months	2.5 years
Kuwait	1	19	8 months	9 years
Libya	1	28	7 months	7 months
Japan	1	19	36 months	6 years
Total	81	21	10 months	5.5 years

Two raters scored the participants' essays. In order to recruit raters for this study, I posted a message on a Facebook page for graduate students and graduates of the MATESOL and SLS programs of Michigan State University. The three criteria for scoring the essays were the following: (1) a master's degree in Teaching English as a Second Language; (2) a minimum of two years of experience teaching ESL academic writing; (3) and prior experience scoring second language learners' writing. I selected two who responded and met the criteria for scoring. The raters were two female native speakers of English in their early thirties. One rater, whom I shall call RM, majored in Linguistics and was a Ph.D. candidate in Second Language Acquisition at the time of data collection. RM taught ESL for three years both to K-8 and college level students. The other rater, RK, had a BA in German and a master's degree in Teaching English as a Second Language. She was a rater for a high stakes ESL standardized test at the time of data collection. RK taught ESL for six years both in K-12 and at the college level.

2.1.2 Procedure

Following the procedures in my previous study (David, under review), the participants took two different writing exams: an impromptu TW exam and a PBTW exam. The order of the exams was counter-balanced and administered within the same week to reduce any effect of instruction. The participants were randomly divided into two groups: one group wrote an impromptu TW exam about obesity and a PBTW exam about gun control; and the other group wrote an impromptu TW exam about gun control and a PBTW exam about obesity. As mentioned above, the order of the TW and PBTW exams was counter balanced. Table 2 outlines the procedure for the administration of the two exams for the two groups.

Table 2
Procedures for the TW and PBTW exams for each group

Group	**Order of the exams/topic**	**Order of the exams/topic**
Group 1	Impromptu TW (obesity)	PBTW exam (gun control)
	PBTW exam (gun control)	Impromptu TW exam (obesity)
Group 2	Impromptu TW exam (gun control)	PBTW exam (obesity)
	PBTW exam (obesity)	Impromptu TW exam (gun control)

The prompts for both topics and both types of exams were exactly the same, with one exception. The participants did not read articles, watch videos, participate in class discussion, or have time to plan their essays for the impromptu TW exam. The procedure followed for each exam is described below.

2.1.3 TW exam

The participants were given a prompt about obesity or gun control, depending on the group to which they were assigned, and they were given 45 minutes to write an essay answering the prompt. Below are the prompts for each topic:

- Obesity is a healthcare concern worldwide, but especially in the United States. Two solutions being proposed are: 1) to tax junk food to discourage people from buying it; and 2) to ban the sales of large sodas in some establishments. Do you believe these solutions would encourage people to reduce their consumption of unhealthy foods? Propose other solutions to the problem in the United States. Be sure to fully develop your essay by including logical supporting ideas, clear explanations, relevant examples, and specific details.
- Gun control continues to be a problem in the United States and in other countries around the world. There are three main views on gun control in the United States: 1) Restrict the sales of guns to people with no criminal background; 2) ban the sales of guns altogether; or 3) allow the sales of guns to anyone. What do you think? Be sure to fully develop your essay by including supporting ideas, clear explanations, relevant examples, and specific details.

The participants were allowed to ask any questions that they had about the language contained in the prompts or questions that clarified what they were expected to do for the writing task. The participants were not allowed to use dictionaries or any electronic devices during the exam.

2.1.4 PBTW exam

I created this exam for another study that I conducted (David, under review). The students watched two short videos (the links can be found in Appendix A) and read an article (the links can be found in Appendix B) about the same topic (obesity or gun control). They were then presented with the prompt and they were given 10 minutes to discuss their ideas in groups. After that, the students had 10 minutes to plan what they would write on a sheet of paper

containing the prompt and blank space for pre-writing. No specific pre-writing technique was elicited and the participants were not allowed to speak to one another during the planning stage. Finally, they had 45 minutes to write their essays. Again, the participants were not allowed to use dictionaries or electronic devices during data collection, but they were allowed to ask questions about the vocabulary or content of the prompt. Below are the prompts for each topic:

- Obesity is a healthcare concern worldwide, but especially in the United States. Two solutions being proposed are: 1) to tax junk food to discourage people from buying it; and 2) to ban the sales of large sodas in some establishments. Do you believe these solutions would encourage people to reduce their consumption of unhealthy foods? Propose other solutions to the problem in the United States. Be sure to fully develop your essay by including logical supporting ideas, clear explanations, relevant examples and specific details. **Use ideas from the videos we watched and the article we read about the topic. Do not forget to give credit to the authors.**
- Gun control continues to be a problem in the United States and in other countries around the world. There are three main views on gun control in the United States: 1) Restrict the sales of guns to people with no criminal background; 2) ban the sales of guns altogether; or 3) allow the sales of guns to anyone. What do you think? Be sure to fully develop your essay by including supporting ideas, clear explanations, relevant examples, and specific details. **Use ideas from the videos we watched and the article we read about the topic. Do not forget to give credit to the authors.**

Table 3, adapted from my earlier study (David, under review), has a detailed description of the procedures that were followed for the PBTW exam about obesity and Table 4 outlines the procedures for the PBTW exam about gun control. The PBTW exam began with a five-minute

introduction to the general topic of the prompt. When the other teachers and I proctored the exam, we asked the participants questions to activate the participants' background knowledge of the topic. After that, we told the participants that they would watch two videos and read one article related to the topic of the essay that they would have to write. We encouraged the participants to take notes while watching the videos and reading the article and said that the students would be allowed to use those notes while planning and writing their essay. After the participants watched the videos and read the article, we introduced the prompt to the students and told them that they would have ten minutes to discuss their ideas about the prompt in groups of four. Before the group discussion began, we asked if the students had any questions about the prompt. While the participants were discussing their ideas, as a proctor we walked around and moved from group to group to help groups who were struggling to share their ideas by asking questions to prompt further discussion. Next, the participants were given a blank sheet of paper with the prompt written on the top of the paper and ten minutes to plan their essay. The participants were told that they could use the planning sheet when writing. Finally, the participants had 45 minutes to write their essay. The proctor had a timer projected in the front of the class so that the participants could keep track of time. Moreover, we instructed the participants to write down how many minutes were remaining when they finished writing their essay in order to measure fluency.

Table 3
Procedures for the PBTW exam for Group 1

Time Frame	**Activity**	**Procedures**
5 minutes	Introducing the topic	1) The proctor introduced the topic by asking students the following questions: What is obesity? What causes obesity? What are the consequences of obesity? How can we encourage people do eat more healthy foods?

Table 3 (Cont'd)

5 minutes	Videos	2) The proctor briefly explained what the videos are about and played them for students. The students were encouraged to take notes while watching the videos.
15 minutes	Reading	3) The students read the article.
10 minutes	Group discussion	4) The proctor read the essay prompt aloud to students and asked if they had any questions about it. 5) The students discussed the essay prompt in groups of four.
10 minutes	Planning	6) Students planned the essay.
45 minutes	Writing	7) Students wrote their essays.

Table 4
Procedures for the PBTW exam for Group 2

Time Frame	**Activity**	**Procedures**
5 minutes	Introducing the topic	1) The proctor introduced the topic by asking students the following questions: What do you know about gun control? What laws does your home country have about gun control? Should anyone be allowed to buy and carry guns?
5 minutes	Videos	2) The proctor briefly explained what the videos are about and played them for students. The students were encouraged to take notes while watching the videos.
15 minutes	Reading	3) The students read the article.
10 minutes	Group discussion	4) The proctor read the essay prompt aloud to students and asked if they had any questions about it. 5) The students discussed the essay prompt in groups of four.
10 minutes	Planning	6) Students planned the essay.
45 minutes	Writing	7) Students wrote their essays.

2.1.5 Rating

Two experienced raters rated the exams using an analytic rubric. Analytic rubrics allow raters to assign scores to individual categories. The total score is a sum of the scores that the test

taker received in each category. The rubric used for this study was developed by Weir (1990) (see Appendix C). One reason why I chose this specific rubric was because Weigle (2002) presented this rubric as a reliable choice when discussing analytic rubrics in her book. The other analytic rubric I could have chosen was Jacobs et al.'s (1981), which is also presented in Weigle's book, but it seemed very wordy in comparison to Weir's. In addition, there is much controversy surrounding Jacobs et al.'s rubric (see Connor-Linton & Polio, 2014; Winke & Lim, 2015). Thus, I decided to opt for Weir's rubric because of its simplicity and because it has not been controversial. The rubric had seven categories: content, organization, cohesion, vocabulary, grammar, punctuation, and spelling. Each category could be assigned a score between zero and three. Before rating the essays, the raters participated in training and norming sessions to become familiar with the rubric and to calibrate their ratings. Training and norming sessions are a crucial part of the rating procedure and many testing experts suggest that raters train before using a new rubric and norm by reading essays together to ensure that their scores are reliable (Carr, 2011; Weigle, 2002). In training sessions, raters read and discuss the rubric together. In norming sessions, raters read sample essays and score them. They then share their scores with other raters and discuss why they assigned such scores. There were two training and norming sessions: one to train and norm for scoring the TW exams and another one to norm for scoring the PBTW exam. At the beginning of the first training and norming session, the raters read and signed consent forms to participate in the study. After each session, the raters had three weeks to score the essays. After they scored all of the essays, they were given a random subset of 20 essays that they had already scored to investigate intra-rater reliability (ten TW essays and ten PBTW essays).

During the first training and norming session, I first explained to the raters the two exams and showed them all of the materials that were used in the exams. Then we read the rubric and discussed how the raters interpreted each of the seven categories of the rubric. We agreed that content was related to how well and clearly the participants answered the prompt. Because the prompts had multiple questions, we decided that students who answered only one or two of the questions from the prompt would not be penalized. For example, some participants discussed only one of the two proposals written in the prompt about obesity. Others did not suggest a proposal at all. These participants' content scores should not suffer because they neglected to answer all of the questions in the prompt. After a long discussion about the difference between organization and cohesion, we agreed that organization was related to the essay as a whole and how it is organized from a macro level perspective (introduction, body, conclusion, thesis statement, topic sentences, and so on), while cohesion was related to how the essay was organized in a micro level or sentence level perspective. We discussed that when scoring vocabulary the raters should pay attention to the complexity of words used in the texts, word choice, and so on. If the students repeat certain words too often and use very simple vocabulary, their vocabulary score should suffer. For grammar, we agreed that the raters should pay attention to verb tense, subject and verb agreement, fragments, and the complexity of the grammatical structures used in the essays. We agreed that the raters should not penalize the participants who made punctuation mistakes around quotations, because investigating source incorporation was not one of the goals of this study. One issue that arose in the training session was that of source integration. One of the raters asked how they should score essays in which test takers did not use sources at all, even though they were instructed to, or essays in which the test takers did not integrate sources very well. After some discussion, we agreed that the test takers should not be

penalized if they failed to use sources or used sources inadequately. This decision was made to ensure that both types of essays were scored based on the same criteria and could, therefore, be compared to one another.

Finally, after norming five essays and having difficulty agreeing with each other's spelling scores, we decided that if an essay had less than five spelling mistakes, the participant would receive a three for spelling; if it had more than five mistakes, the participant would receive a two; and if the essay had spelling mistakes in every other sentence, the participant would receive a one. If a student repeated the name spelling mistake throughout the essay, the mistake should be counted as one. After reading the rubric, the raters read and scored six essays. They read one essay quietly and assigned scores for each category without consulting one another. Next, they shared their scores and discussed why they assigned the scores for each category. This procedure was repeated for each of the six essays that they scored. In the second norming session, the raters shared problems that they were having with the rubric and potential ways to solve them and then they scored six essays following the same procedures they did in the first norming session. The sessions were audio-recorded with an Olympus digital voice recorder to gather information about the raters' impressions about the two exams.

2.1.6 Post-writing questionnaire

All learners completed a post-writing questionnaire that included multiple-choice and short-answer questions about the two exams (see Appendix D). The questions were related to what they liked and disliked about the exams, what they thought was difficult or easy about each exam, what they thought about the article, videos, group discussion, and so on. The participants took approximately ten minutes to answer the questionnaire.

2.1.7 Semi-structured interviews

I randomly selected four students from the two 220 classes that I taught to participate in a semi-structured interview in groups to discuss their attitudes about the two exams and obtain more detailed information about their perceptions of the two exams. The teachers of the other four classes in which data were collected asked for volunteers or selected students to participate in the semi-structured interviews for me. In other words, some students volunteered for the interviews in some classes and other students were chosen by their teachers for the interviews in other classes. There were six interviews, and they each lasted approximately 15 to 20 minutes. In the first and second interviews, there were four participants (ID07, ID09, ID13, and ID16 and ID21, ID22, ID24, and ID25); in the third and fourth interviews there were three participants (ID42, ID43, and ID45 and ID49, ID50, and ID51); and there were two participants in the fifth and sixth interviews (ID59 and ID60 and ID73 and ID74). As I was introducing myself to the participants before data collection began, I identified myself to them as Brazilian and a graduate student. At the time of data collection, there were multiple Brazilian students enrolled at the English Language Center as part of a government program called *Ciências Sem Fronteiras* (Science Without Borders). All of the participants in the third and sixth interview groups were from Brazil and before the interview began I asked them if they wanted to be interviewed in English or in Portuguese, since the latter is my native language. They expressed that they preferred that the interview be conducted in Portuguese. For these interviews, I transcribed them in Portuguese and provided English translations in italics. As I mentioned above, I collected data in two ESL 220 classes for which I was the instructor. Although there were many students from Brazil in those two classes, I did not give them the option to conduct the interview in Portuguese.

I was their teacher at the time of data collection and did not want to encourage them to speak Portuguese to me in class. Therefore, I never spoke to them in Portuguese at all. There were four different ways in which I identified myself to the students in this study and that could have influenced the way that the participants interacted with me during the interviews: Brazilian, student, researcher, and teacher. I was a fellow Brazilian to some participants; a fellow student to others; a researcher to all; and a teacher to some. I will discuss these identities and how they could have affected my interactions with the participants when I discuss the students' perceptions of the exam. The raters also participated in a semi-structured interview to discuss their thoughts about the exams and the rubric. The raters, however, unlike the students, were interviewed separately. The interviews were audio recorded with an Olympus digital voice recorder. The questions for the semi-structured interviews are in Appendix E.

2.2 Analysis

To answer the first question, that is, whether the test takers' scores in an impromptu TW exam correlate with their scores on a PBTW exam when raters use a holistic and an analytic rubric, I entered all scores into IBM SPSS version 21 and ran a Spearman correlation due to the fact that the scores each participant received should be considered a discrete (ordinal) variable. If the data were interval (scale data), then a Pearson correlation should have been used, according to Field (2009). I ran a paired samples *t*-test with the participants' scores to determine whether there were any significant differences between the scores in the two exams to answer the second research question.

To examine how students' writing across exams differed in terms of grammatical accuracy, a research assistant first counted the number of clauses and then the number of error-

free clauses in the TW and PBTW essays, following Ellis and Yuan (2004) and Wigglesworth and Storch (2009). Before she analyzed the essays for accuracy, the research assistant and I met to read Polio's (1997) guidelines for clauses and errors (see Appendix F). We then read and analyzed two essays together, discussing any issues or disagreements we had. We divided the essays into clauses and then discussed whether each clause contained an error. As we did so, we decided to add the following guidelines to Polio's:

a. Spelling errors that result in a completely different word are counted as word choice errors;
b. If there is a comma splice, the clause preceding the comma splice is the one to which the error will be added;
c. In the case of the PBTW exam, do not include quotations as clauses.
d. Accept "the" as a misspelling of "they".

We decided not to include quotations in the clause count for obvious reasons. Quotations were grammatically accurate almost 100% of the time and the accuracy was not a result of the participants' interlanguage, because they did not write the sentence. We found that many of the Arabic speaking participants used "the" instead of "they" and we agreed that this was a spelling mistake and therefore should not be counted as an error, as determined by Polio (1997). After we analyzed the two essays together, the research assistant started her analysis of accuracy in the two exams and I analyzed 10% of the essays (or 17 essays) to check for inter-rater reliability using Cronbach's alpha. I then divided the number of error-free clauses by the total number of clauses in each essay and used the percentage of error-free clauses for the analysis.

I used two programs to analyze three aspects of lexical complexity: RANGE, developed by Nation (2005), and the Lexical Complexity Analyzer, developed by Lu (2010). The two

aspects of lexical complexity that I investigated were lexical sophistication, lexical density, and lexical variation. Before entering the essays into the two programs, I typed them and saved them as .txt files, which is the type of file that both programs have to use. As described above, RANGE compares texts to three different word lists: the most common 1,000 words, the second most common 1,000 words, and a list of university words. I used RANGE to investigate lexical sophistication, as Fritz and Luegg (2013) and Kormos (2011) did in their studies. The Lexical Complexity Analyzer generates 25 different measures of lexical complexity, which include lexical density and variation. However, I did not use the 25 measures in my analysis of lexical complexity. Following Kormos (2013) and Lu (2010), I used Malvern and Richards' (1997) D-Measure as one measure of lexical variation. The other measures of lexical variation generated by the Lexical Complexity Analyzer were: lexical word variation, verb variation 1, noun variation, adjective variation, and adverb variation.

To investigate syntactic complexity I ran the participants' essays in the L2 Syntactic Complexity Analyzer (Lu, 2010). The L2 Syntactic Complexity Analyzer generates 14 different syntactic complexity measures: mean length of sentence, mean length of T-unit, mean length of clause, clause per sentence, verb phrase per T-unit, clause per T-unit, dependent clause per clause, dependent clause per T-unit, T-unit per sentence, complex T-unit ratio, coordinate phrase per T-unit, coordinate phrase per clause, complex nominal per T-unit, and complex nominal per clause. Lu (2010) correlated the complexity scores generated by the program with scores generated by two human raters and found the correlations to be high, ranging from .845 to 1. The author concluded that the L2 Syntactic Complexity Analyzer is indeed a reliable measurement of complexity. In order to investigate fluency, I used Microsoft Word to get word counts for each essay. In addition, I asked the participants to write down how many minutes were left before they

turned in their essays to measure how many words per minute they wrote. Following Kellogg (1988), I used three different numbers to compare fluency in the TW and PBTW exams: total number of words per essay, total time writing, and total number of words per minute. Although these measures of fluency may not be the most valid (Abdel Latif, 2012), as described above, the lack of consensus and research on which measures are more valid left me no choice but to choose ones that researchers have often used.

In order to determine the intra-rater reliability for each rubric, I obtained a Cronbach's alpha coefficient for the twenty essays that the raters rated twice. Similarly, I obtained a Cronbach's alpha coefficient based on the total scores that each rater assigned to each essay to determine inter-rater reliability. Moreover, I also obtained Cronbach's alpha coefficients for the scores in each of the seven categories in the analytic rubric. Spearman correlations were also used to examine intra and inter-rater reliability.

To answer the fifth research question, which asks about students' perceptions of the two different types of writing exams, I inputted the participants' responses from the post-writing questionnaire into SPSS and obtained a count for their answers in the multiple-choice questions. This set of data was more quantitative in nature and was converted into percentages. For example, for the first question, which asked the participants which exam they thought was easier, I counted how many participants circled each multiple choice option (the TW exam, the PBTW exam, or they were both equally easy or difficult) and then calculated the percentage for each of them. This procedure was followed for all of the multiple-choice questions. However, there were also questions that were open ended and usually followed up on a multiple-choice question. Before analyzing the data, I typed all of the participants' responses on a Microsoft Word document. This set of data was more qualitative and therefore I followed Baralt's (2012)

guidelines for coding qualitative data. According to her, the first step to analyzing qualitative data is to read through the data the first time and start thinking of ways that the data could be coded. This is called open coding, which is "the process of assigning a code to represent a concept shown in the data" (Baralt, 2012, p. 230). Instead of denoting the codes myself, I chose to use a more emergent approach, what Baralt called *in vivo* code. *In vivo* codes arise from the data and are not assigned by the researcher. As I noticed codes emerging from the data, I highlighted any text that seemed to discuss the same code and then created a name for the code. For example, many wrote that the videos, articles, and discussions helped them to think of ideas to use in their essay. As I noticed that, I re-read the text and highlighted anything that the participants wrote that were related to that topic. Baralt suggests that the researcher go through more than one iteration with the data. She explained that after coding the data, the researcher should read through the data again and refine the themes to better understand to what they relate. Next, the researcher has to re-read all of the data coded under one theme and compare it to ensure that everything should indeed be coded under that particular theme. For example, many participants wrote about time. However, after re-reading the data, I noticed that some participants wrote about time to write the essay and others wrote about time to plan the essays. While both ideas are related to time, one is related to planning time, whereas the other is related to time to write, which are two different types of time. According to Baralt, I can choose to separate these themes into two or create one overarching theme with subcategories to combine them. I chose the first option and coded some of the responses as planning time and others as time. Finally, the researcher has to interpret the data. In Baralt's words, this is "the researcher's opportunity to explain how the research questions were answered" (2012, p. 234), which is what I do in chapter 4.

The manner in which I approached the interview data with the test takers was somewhat different from the manner in which I dealt with the written qualitative data from the post-writing questionnaires. All of the participants answered the post-writing questionnaire, whereas only 18 participants participated in the semi-structured interviews. Since the interviews served as complement to the responses in the post-writing questionnaire, I carefully listened to the interviews and transcribed only quotes that described the themes which emerged in the questionnaire. As Baralt suggested, and as I did with the written questionnaire data, I re-listened to the interviews multiple times and merged or separated codes as needed.

Following Baralt's suggestions for coding qualitative data described above, I started by carefully listening to and transcribing the interviews with the raters to answer the last research question, which asked about the rater's perspectives about the exams. Again, I listened the first time to begin the open coding of the data and then I re-listened multiple times to fine tune the codes and create themes. I then listened to the two norming sessions and transcribed quotes that belonged to the same themes that emerged in the interviews. As I did with the interviews with the test takers, the norming sessions served as a supplement to the themes that arose in the interviews.

CHAPTER 3: QUANTITATIVE RESULTS

First, I describe the results of the first four research questions, all of which are quantitative questions. In chapter 4, I report on the results of the last two research questions, which are qualitative questions. Below are the results for each quantitative research question.

3.1 RQ1: Do test takers' scores in an impromptu TW exam correlate with their scores in a PBTW exam?

In order to determine whether the participants' scores in the two exams correlated, I calculated the average score for each category and for each participant by adding the scores that the two raters assigned and dividing them by two. For example, if one rater assigned a participant a score of 1 and the other rater assigned a score of 2 for spelling, the average score would be 1.5. Most of the scores that the participants assigned differed by one point (one rater assigned a 1 and the other a 2) or were exactly the same and only 4% of them differed by two points (one rater assigned a 1 and the other assigned a 3). After that, I entered all of the scores in IBM SPSS and ran a Spearman correlation. Table 5 shows the descriptive statistics for the participants' scores in the TW and in the PBTW exams.

Table 5
Descriptive statistics: Average scores

	N	Minimum	Maximum	Mean	Std. Deviation
TW exam	81	9.5	19	14.88	2.17
PBTW exam	81	10.5	20.5	14.95	2.35

The results of the Spearman correlation revealed that the average scores that the participants received in the TW exam and in the PBTW exam only correlated moderately, according to the levels set by Cohen (1992). Cohen suggests that a small correlation is

approximately .10, a medium correlation is .30, and a large correlation is .50. Table 6 shows the results of the Spearman correlation.

Table 6
Spearman's correlation: Average scores

	Spearman's rho	**Sig.**
TW and PBTW exams	.391	.000

In addition to analyzing the correlation between the average scores in the TW and PBTW exams, I also investigated how each score in the different categories of the analytic rubric correlated. The descriptive statistics for each score in the analytic rubric are in Table 7.

Table 7
Descriptive statistics: Average analytic scores

Scores	**N**	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
Content TW	81	1	3	2.19	.52
Content PBTW	81	1.5	3	2.40	.52
Organization TW	81	1	3	1.88	.50
Organization PBTW	81	1	3	1.87	.54
Cohesion TW	81	1	3	2.17	.52
Cohesion PBTW	81	1	3	2.14	.50
Vocabulary TW	81	1	3	2.14	.49
Vocabulary PBTW	81	1	3	2.06	.44
Grammar TW	81	1	3	1.90	.36
Grammar PBTW	81	1	2.5	1.80	.36
Punctuation TW	81	1.5	3	2.37	.42
Punctuation PBTW	81	1.5	3.2	2.56	.38
Spelling TW	81	1	3	2.29	.61
Spelling PBTW	81	1	3	2.03	.58

The results of the Spearman correlations for the analytic scores revealed moderate correlations between the scores the participants received for vocabulary, punctuation, and spelling, and weak correlations between the scores that the participants received for content, organization, cohesion, and grammar. The results of the correlations for the analytic scores are in Table 8.

Table 8
Spearman correlations: Average analytic scores

Scores	Spearman's rho	Sig
Content	.218	.050
Organization	.246	.027
Cohesion	.178	.112
Vocabulary	.302	.006
Grammar	.221	.048
Punctuation	.380	.000
Spelling	.355	.001

3.2 RQ2: Do test takers' scores in an impromptu TW exam differ significantly from their scores in a PBTW exam?

In order to investigate whether the participants' scores differ in the TW and PBTW exams, I used the test takers' average scores. As I mentioned before, I calculated the average scores by adding the scores the two raters assigned to each participant and dividing them by two. I entered the scores in SPSS and ran a paired-samples *t* test with the average scores. Below are the results of the *t* test.

Table 9
T test: Average scores

	Mean	Std. Deviation	Std. Error Mean	*t*	Sig.	Effect size
TW and PBTW exams	.07	2.48	.27	-.268	.789	.01

The *t* test did not reveal a significant difference between the scores that the participants received in the TW and PBTW exams. In order to examine the scores in more detail, I ran *t* tests for each of the average analytic scores that the participants received. Running too many *t* tests increases the chances that Type I error will occur. Type I error is a false positive; that is, Type I error is when the results show a significant difference when there was none (Larson-Hall, 2009). To reduce the risk of type I error, Larson-Hall suggested doing a Bonferroni adjustment, which is "decreasing the acceptable alpha rate depending on how many *t* tests are done" (2009, p. 252).

The author went on to explain that "the Type I error rate is less than or equal to the number of comparisons done, multiplied by the chosen alpha level" (p. 252). To fix this problem, according to her, the researcher should divide the alpha level by the number of comparisons being done with the test (Larson-Hall, 2009). The desired alpha level for the purpose of my research is .05 and the number of comparisons done for the analytic scores is 7 (7 different analytic scores): .05 divided by 7 is .007. Anything above .007 will not be considered statistical. Table 10 shows the results of the *t* tests for the average analytic scores.

Table 10
T tests: Average analytic scores

Scores	**Mean**	**Std. Deviation**	**Std. Error Mean**	***t***	**Sig**	**Effect size**
Content	.20	.66	.07	2.859	.005	.19
Organization	.01	.66	.07	.133	.867	.01
Cohesion	.03	.65	.07	.107	.613	.03
Vocabulary	.08	.55	.06	.042	.198	.08
Grammar	.09	.45	.05	.000	.052	.13
Punctuation	.19	.44	.04	.293	.000	.23
Spelling	.25	.67	.07	-.103	.001	.20

The results of the *t* tests revealed that there were significant differences in the scores that the participants received for the following categories in the analytic rubric: content, punctuation, and spelling. As can be seen in Table 10, the test takers scored significantly higher for content and punctuation in the PBTW exam and they scored significantly higher for spelling in the TW exam.

3.3 RQ3: How do students' writing across exams differ in terms of:

3.3.1 Accuracy

Following Ellis and Yuan (2004) and Wigglesworth and Storch (2009), accuracy was measured in terms of error-free clauses. After the research assistant counted the number of clauses and error free-clauses in both TW and PBTW exam, I entered the percentage of error-free clauses for each essay in IBM SPSS. The descriptive statistics for the percentage of error-free clauses in the TW and PBTW exams are in Table 11.

Table 11
Descriptive statistics: Percentage of error-free clauses

	N	Minimum	Maximum	Mean	Std. Deviation
TW exam	81	0	62.06	31.85	14.43
PBTW exam	81	2.77	76.47	33.08	14.09

In order to determine whether the accuracy scores differed in the two exams, I ran a paired-samples *t* test. The results of the *t* tests are in Table 12. The results of the *t* test revealed that there were no significant differences between the accuracy scores in the TW exam and the PBTW exam. In order to measure rater reliability, I analyzed accuracy in approximately 10% of the essays. The Cronbach's alpha coefficient for inter-rater reliability obtained was .842, which is high and on par with essay-test reliabilities in large-scale and high stakes tests like the TOEFL (www.toefl.org).

Table 12
T test: Accuracy

	Mean	Std. Deviation	Std. Error Mean	*t*	Sig.	Effect size
Error-free clauses	1.23	16.37	1.81	-.677	.501	.04

3.3.2 Lexical complexity

I used the programs RANGE (Nation, 2005) and the Lexical Complexity Analyzer (Lu, 2010) to analyze the 162 essays for lexical complexity in three different dimensions: lexical sophistication, lexical density, and lexical variation. As I mentioned previously, RANGE compared the essays to three different word lists (the first and second most common 1,000 words and university words) to investigate lexical sophistication. The descriptive statistics for lexical sophistication are in Table 13.

Table 13
Descriptive statistics: Lexical sophistication in TW exams and PBTW exams

Word lists	**N**	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
Word list 1 TW	81	155	472	263.71	70.12
Word list 1 PBTW	81	154	463	290.18	74.82
Word list 2 TW	81	7	49	24.18	8.50
Word list 2 PBTW	81	8	54	28.64	10.97
Word list 3 TW	81	0	30	7.95	5.39
Word list 3 PBTW	81	0	25	10.04	5.80

In order to determine whether the participants' essays differed significantly in terms of lexical sophistication, I ran three *t* tests and set the alpha level to .016 to avoid Type I error, as suggested by Larson-Hall (2009). The desired alpha level for the purpose of my research is .05 and the number of comparisons done for syntactic complexity is three (three different word lists): .05 divided by 3 is .016. Anything above .016 will not be considered statistical. Table 14 shows the results of the *t* tests for the three word lists.

Table 14
T test: Lexical sophistication

	Mean	**Std. Deviation**	**Std. Error Mean**	***t***	**Sig.**	**Effect size**
Word list 1	26.46	61.93	6.88	3.846	.000	.17
Word list 2	4.45	13.51	1.50	2.968	.004	.22
Word list 3	2.09	5.70	.63	3.310	.001	.18

The results of the *t* test revealed a significant difference between the PBTW and TW exams for all three word lists. The participants used significantly more words in the first 1,000 most common words list, in the second 1,000 most common words list, and in the university words list in the PBTW exam.

The Lexical Complexity Analyzer generated 25 different measures of lexical density and lexical variation. For the purposes of this study, I only use the D-measure for lexical density and the following measures to investigate lexical variation: lexical word variation, verb variation 1, noun variation, adjective variation, and adverb variation. The descriptive statistics for these measures of lexical density and variation are shown in Table 15.

Table 15
Descriptive statistics: Lexical Density and Lexical Variation

Lexical measures	**N**	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
Word Type TW	81	90	210	129.13	25.21
Word Type PBTW	81	87	215	146.80	28.93
Lexical Density TW	81	.46	.64	.53	.03
Lexical Density PBTW	81	.48	.64	.53	.03
Lexical Variation TW	81	.37	.79	.56	.09
Lexical Variation PBTW	81	.38	.79	.56	.08
Verb Variation I TW	81	.37	1	.66	.12
Verb Variation I PBTW	81	.37	.83	.65	.10
Noun Variation TW	81	.32	.77	.49	.10
Noun Variation PBTW	81	.35	.75	.52	.08
Adjective Variation TW	81	.06	.18	.10	.02
Adjective Variation PBTW	81	.05	.21	.10	.03
Adverb Variation TW	81	.02	.09	.05	.01
Adverb Variation PBTW	81	.01	.11	.05	.01

Once again, I ran *t* tests to determine whether there were significant differences in the two exams regarding lexical density and lexical variation. The desired alpha level for the purpose of my research is .05 and the number of comparisons done for lexical density and lexical variation is seven: .05 divided by 7 is .007. Anything above .007 will not be considered statistical. Table 16 shows the results of the *t* tests for lexical density and lexical variation.

Table 16
T tests: Lexical Density and Lexical Variation

Lexical measures	**Mean**	**Std. Deviation**	**Std. Error Mean**	***t***	**Sig.**	**Effect size**
Word Type	17.66	22.52	2.50	-7.059	.000	.30
Lexical Density	.002	.04	.005	-.562	.576	.03
Lexical Variation	.009	.07	.008	-1.126	.264	.05
Verb Variation I	.01	.11	.01	.772	.442	.04
Noun Variation	.02	.08	.009	-2.994	.004	.14
Adjective Variation	.003	.03	.003	1.062	.291	.06
Adverb Variation	.001	.02	.002	-.815	.417	.04

The results of the *t* test revealed that there was a significant difference in word type and noun variation between the TW and PBTW exams. The participants used significantly more word types and a wider variation of nouns in the PBTW exam when compared to the TW exam.

3.3.3 Syntactic complexity

As mentioned above, in order to measure syntactic complexity, I used the L2 Syntactic Analyzer developed by Lu (2010). The L2 Syntactic Complexity Analyzer generated 14 different syntactic complexity measures: mean length of sentence (MLS), mean length of T-unit (MLT), mean length of clause (MLC), clause per sentence (C/S), verb phrase per T-unit (VP/T), clause per T-unit (C/T), dependent clause per clause (DC/C), dependent clause per T-unit (DP/T), T-unit per sentence (T/S), complex T-unit ratio (CT/T), coordinate phrase per T-unit (CP/T), coordinate phrase per clause (CP/C), complex nominal per T-unit (CN/T), and complex nominal per clause (CN/C). Table 17 shows the descriptive statistics for the 14 measures of syntactic complexity for the PBTW and TW exams.

Table 17
Descriptive statistics: Syntactic complexity

Lexical Complexity Measures	N	Minimum	Maximum	Mean	Std. Deviation
MLS PBTW	81	9.66	48.25	19.27	5.60
MLS TW	81	10.88	39.83	19.15	4.91
MLT PBTW	81	9.32	38.6	16.56	4.13
MLT TW	81	11.08	34.14	16.40	3.75
MLC PBTW	81	6.26	13.18	9.02	1.47
MLC TW	81	6.23	14.29	8.90	1.64
C/S PBTW	81	1.14	6.00	2.17	.73
C/S TW	81	1.13	4.11	2.19	.60
VP/T PBTW	81	1.46	5.1	2.50	.62
VP/T TW	81	1.70	4.85	2.52	.53
C/T PBTW	81	1.10	4.80	1.85	.50
C/T TW	81	1.13	2.88	1.86	.40
DC/C PBTW	81	.19	.60	.37	.08
DC/C TW	81	.11	.59	.36	.09
DC/T PBTW	81	.25	2.08	.72	.32
DC/T TW	81	.13	1.69	.71	.32
T/S PBTW	81	.96	1.50	1.15	.11
T/S TW	81	.96	1.61	1.16	.13
CT/T PBTW	81	.21	.80	.49	.14
CT/T TW	81	.13	.84	.49	.16
CP/T PBTW	81	0	1.10	.32	.18
CP/T TW	81	0	1.14	.32	.18
CP/C PBTW	81	0	.50	.17	.09
CP/C TW	81	0	.70	.17	.10
CN/T PBTW	81	.18	4.10	1.82	.63
CN/T TW	81	.82	3.88	1.85	.61
CN/C PBTW	81	.52	1.97	1.00	.28
CN/C TW	81	.48	1.84	1.00	.29

In order to determine whether the participants' essays differed in the measures of syntactic complexity above, I ran paired samples *t* tests on SPSS. As I mentioned before, running too many *t* tests increases the chances that Type I error will occur. In order to avoid Type I error, I adjusted the alpha level as suggested by Larson-Hall (2009). The desired alpha level for the purpose of my research is .05 and the number of comparisons done for syntactic complexity is 14 (14 different syntactic complexity measures): .05 divided by 14 is .0035. Anything above .0035

will not be considered statistical. Table 18 shows the results of the *t* tests for the 14 different measures of syntactic complexity.

Table 18
T tests: Syntactic complexity

Lexical Complexity Measures	**Mean**	**Std. Deviation**	**Std. Error Mean**	***t***	**Sig**	**Effect size**
MLS	.119	4.64	.51	.231	.818	.01
MLT	.154	4.01	.44	.347	.729	.01
MLC	.123	1.82	.20	.608	.545	.03
C/S	-.017	.69	.76	-.233	.816	.01
VP/T	-.016	.60	.67	-.241	.810	.01
C/T	-.009	.52	.58	-.163	.871	.01
DC/C	.004	.11	.01	.321	.749	.02
DC/T	.003	.40	.04	.067	.947	.004
T/S	-.007	.14	.01	-.437	.663	.02
CT/T	-.011	.18	.02	-.570	.570	.03
CP/T	-.000	.20	.02	-.005	.996	.0002
CP/C	-.003	.12	.01	-.241	.810	.01
CN/T	-.031	.71	.07	-.402	.689	.02
CN/C	.002	.33	.03	.060	.952	.003

As you can see from Table 17, there were no significant differences between any of the 14 different syntactic complexity measures in the PBTW and TW exams.

3.3.4 Fluency

To measure fluency, I asked the participants to write down on the test sheet how many minutes were left on the timer when they finished writing their essays. In order to measure fluency, I entered the number of words the participants wrote on SPSS, the time it took them to write their essays, and the amount of words they wrote per minute. As mentioned previously, the participants had 45 minutes to write both essays. However, they spent more time on task when they took the PBTW exam, because they watched two videos, read one article, had a group discussion, and planned their essays before they started writing. The mean number of words that

the students wrote in the PBTW exam was 353.9 and 315.6 in the TW exam. Table 19 describes the descriptive statistics for the number of words in the TW and PBTW exams.

Table 19
Descriptive statistics: Number of words

	N	Minimum	Maximum	Mean	Std. Deviation
TW exam	81	194	550	315.6	88.1
PBTW exam	81	195	536	353.9	77.8

A *t* test revealed a significant difference for the total number of words between the TW exam and the PBTW exam. The participants wrote significantly longer essays in the PBTW exam. Table 20 has the results of the *t* test.

Table 20
T test: Number of words

	Mean	Std. Deviation	Std. Error Mean	*t*	Sig.	Effect size
TW and PBTW exams	38.32	70.017	7.78	4.926	.000	.22

Furthermore, the mean number of minutes that students wrote for the PBTW exam was 41.3 minutes and 40.6 minutes for the TW exam. Table 21 shows the descriptive statistics for the number of minutes that students wrote for each exam. A *t* test revealed no significant differences between the number of minutes that the participants wrote for the two exams (*t*(1.525), *p*(.131), *r*(.54)).

Table 21
Descriptive statistics: Number of minutes

	N	Minimum	Maximum	Mean	Std. Deviation
TW exam	81	30	45	40.6	3.9
PBTW exam	81	22	45	41.3	4.9

The participants wrote an average of 8.6 words per minute in the PBTW exam and 7.9 words per minute in the TW exam. Table 22 shows the descriptive statistics for the amount of words that the participants wrote per minute.

Table 22
Descriptive statistics: Words per minute

	N	Minimum	Maximum	Mean	Std. Deviation
TW exam	81	4.7	14.8	7.9	2.2
PBTW exam	81	4.3	14.8	8.6	2.3

A *t* test revealed that the participants wrote significantly more words per minute in the PBTW exam. Table 23 describes the results of the *t* test.

Table 23
T test: Words per minute

	Mean	Std. Deviation	Std. Error Mean	*t*	Sig.	Effect size
TW and PBTW exams	.694	1.88	.20	3.323	.001	.15

3.4 RQ4: What are the intra and inter-rater reliability coefficients for each exam?

To measure intra and inter-rater reliability, I calculated Cronbach's alpha, as suggested by Howell (2002). I also performed Spearman's correlations as a second measure of reliability. The inter-rater reliability coefficient obtained for the scores for the TW exams was .728 and for the PBTW exam was .643. Since the rubric that the raters used was analytic, I decided to investigate the inter-rater reliability for each of the categories of the rubric. Table 24 shows the results Cronbach's analysis and Table 25 shows the results of the correlations. The highest Cronbach's alpha and correlation coefficients were for spelling for both the TW and the PBTW exam. The highest Cronbach's alpha and correlation coefficients for the TW exam were for cohesion, content, and organization. The lowest Cronbach's alpha and correlation coefficients for the TW exam were for grammar, followed by punctuation and vocabulary. The highest

Cronbach's alpha and correlation coefficients for the PBTW exam after spelling were for content and organization, whereas the lowest were vocabulary, followed by punctuation, grammar, and cohesion. The Spearman's rho for the TW exam was .618 and for the PBTW exam was .555.

Table 24
Inter-rater reliability: Analytic scores

Scores	**Cronbach's alpha for TW**	**Cronbach's alpha for PBTW**
Content	.512	.595
Organization	.482	.485
Cohesion	.557	.416
Vocabulary	.479	.242
Grammar	.012	.352
Punctuation	.225	.301
Spelling	.845	.755

Table 25
Correlation matrix for raters' scores

Scores	**Spearman's rho for TW**	**Spearman's rho for PBTW**
Content	.375	.464
Organization	.303	.325
Cohesion	.406	.276
Vocabulary	.302	.131
Grammar	.015	.205
Punctuation	.132	.196
Spelling	.723	.624
Total score	.618	.555

After the raters scored all 162 essays they were given a random selection of 20 essays to score again to measure intra-rater reliability. Ten of these essays were TW exam essays and the other ten were PBTW exam essays. The Cronbach's alpha obtained for RM for all of twenty of the exams was .696 and for RK it was .794. The Cronbach's alpha obtained for RM for the TW exams was .862 and for the PBTW exams it was .352. For RK, the Cronbach's alpha obtained

for the PBTW exam was .552 and for the TW exam it was .645. Below are the correlation matrixes for each category in the analytic rubric for each of the raters.

Table 26
Correlation matrix for RM

	TW Correlations	PBTW Correlations
Content	.544	.375
Organization	.000	.167
Cohesion	.559	.151
Vocabulary	.408	.283
Grammar	.559	.111
Punctuation	.408	.327
Spelling	.808	.377
Total score	.857	.000

Table 27
Correlation matrix for RK

	TW Correlations	PBTW Correlations
Content	.749	.717
Organization	.818	.574
Cohesion	.820	.791
Vocabulary	.745	.856
Grammar	.244	.258
Punctuation	.218	.645
Spelling	.249	.167
Total score	.611	.777

3.5 Summary of the quantitative findings

The first research question asked how the scores in the TW and PBTW exams correlated. The result of the Spearman's correlation was .391, which is a moderate correlation. In addition to correlating the total scores of both exams, I also correlated the scores that the participants received for each category in the analytic rubric. The results for the Spearman's correlations for each category were: content .218; organization .246; cohesion .178; vocabulary .302; grammar .221; punctuation .380; and spelling .355. The correlations for content, organization, cohesion, vocabulary, and grammar were all weak, according to the levels set by Cohen (1992), and the

correlations for vocabulary, punctuation, and spelling were moderate. The second research question asked whether there were significant differences between the scores that the participants received in the two exams. The results of the *t* test revealed that there were no significant differences between the scores in the two exams. However, I also decided to investigate whether there were significant differences between the scores that the participants received for the different categories in the analytic rubric. As mentioned before, because of the increased chance of Type I error caused by performing multiple *t* tests, I followed Larson-Hall's (2009) suggestions and lowered the alpha level for this analysis. The results of the *t* test showed significant differences in the following categories: content, punctuation, and spelling. The participants received significantly higher scores for content and punctuation in the PBTW exam and significantly higher scores for spelling in the TW exam.

The third research question investigated how the TW and PBTW essays differed in terms of accuracy, lexical complexity, syntactic complexity, and fluency. In order to measure accuracy, a research assistant counted the number of clauses and then the number of error-free clauses per essay. After that, I divided the number of error-free clauses by the number of total clauses to obtain a percentage score. I ran a paired-samples *t* test to investigate whether the TW exam and the PBTW exam differed in terms of accuracy and the results revealed no significant differences in the accuracy scores for the two exams. I investigated three different levels of lexical complexity: lexical sophistication, lexical density, and lexical variation. To investigate lexical sophistication, I used RANGE, developed by Nation (2005), which compared the essays using three different corpora: the list of the most common 1,000 words; the list of the second most common 1,000 words; and the list of university words. The results of the *t* test showed that the participants used significantly more words from all three word lists in the PBTW exam. I used

the Lexical Complexity Analyzer (Lu, 2012) to analyze lexical density and lexical variation. This program analyzes essays for 25 different measures of lexical complexity. For the purpose of this study, I only used seven of these measures: lexical density, to measure lexical density, and six measures of lexical variation. The six measures that I used are the following: word type, lexical, verb, noun, adjective, and adverb variation. The reason for choosing six measures, as explained before, was because many of these measures assess the same element of lexical complexity and to reduce the amount of statistical tests and therefore reduce the risk of Type I error. Once again, to reduce the chance of Type I error, I lowered the alpha level for the *t* test. The results of the *t* tests revealed a significant difference only for two measures of lexical variation: word type and noun variation. The participants used significantly more different word types and types of nouns in the PBTW exam. To measure syntactic complexity, I used Lu's (2010) L2 Syntactic Complexity Analyzer, which generated 14 different measures of syntactic complexity. The results of the *t* test, however, revealed no significant differences for the TW and PBTW exams in any of these measures. Finally, I used three different measures of fluency: number of words per essay, number of minutes writing the essay, and total number of words per minute. The results of the *t* test revealed significant differences for the total number of words per essay and the total number of words per minute. The participants wrote significantly longer essays and significantly more words per minute in the PBTW exam.

The fourth, and last quantitative question, asked about the intra- and inter-rater reliability for the two exams. I used two different measures of reliability: the Cronbach's alpha coefficient and the Spearman's correlation. The inter-rater reliability coefficient for the TW exam was .728 and the coefficient for the PBTW exam was .643. According to Carr (2011), a satisfactory coefficient for low stakes exams should be above .700. The Spearman's rho for the TW exam

was .618 and for the PBTW exam the Spearman's rho was .555, both of which are strong correlations. The Cronbach's alpha coefficient for the 20 essays that the raters scored twice were the following: .696 for RM and .794 for RK. In addition to analyzing the Cronbach's alpha coefficient for all 20 essays, I also investigated the coefficient for each of the two exams. The coefficient for the TW essays were: .862 for RM and .552 for RK. For the PBTW exam, the Cronbach's alpha coefficients were: .352 for RM and .645 for RK. Both raters were more consistent when scoring TW exams. I now present information about the 18 participants who participated in the semi-structured interviews.

CHAPTER 4: QUALITATIVE RESULTS

In this chapter, I describe the results for the two qualitative questions, that is, what are the students' (research question 5) and raters' (research question 6) perceptions of the two exams? First, I give more detailed information about the students who participated in the semi-structured interviews. Then, I describe the results of the post-writing questionnaire that students answered about their perceptions of the two exams. Next, I report on the main themes that emerged in the semi-structured interviews with the students. I end this chapter by describing the main themes that emerged in the norming sessions and semi-structured interviews with the raters.

4.1 Participants from the semi-structured interviews

Of the 81 test takers who participated in this study, 18 were randomly selected or volunteered to participate in the semi-structured interviews. The students were grouped according to the 220 or 221 class in which they were enrolled. Students from the same class were interviewed together. The interview groups ranged from two to four participants. The reason for this was because some students did not come to the interview sessions. It may not be ideal to have differing number of students per interview. In a group of four people, the participants might be more willing to share their true thoughts because everybody's opinions differ, while in a smaller group, the participants might be shyer and less willing to share what they really think. One participant might be very talkative and the shier participant might stay quiet or agree with the more dominant talker. These participants were: ID07, ID09, ID13, ID16, ID21, ID22, ID24, ID25, ID42, ID43, ID45, ID50, ID51, ID59, ID60, ID73, and ID74. Table 28 illustrates the groups in which the participants were interviewed.

Table 28
Interview groups

Groups	Participants in each group
Group 1	ID07, ID09, ID13, and ID16
Group 2	ID21, ID22, ID24, and ID25
Group 3	ID42, ID43, and ID45
Group 4	ID49, ID50, and ID51
Group 5	ID59 and ID60
Group 6	ID73 and ID74

Table 29 describes in detail the background information collected from each of the participants who participated in the semi-structured interviews. LoR stands for length of residence in the United States, in months; FI is the number of months the participants have had formal instruction in English, at home or in the United States; and WA stands for writing ability, for which the participants gave self-ratings on a scale of 1 to 5.

Table 29
The participants

Participant	Age	Gender	Nationality	LoR	FI	Major	WA
ID07	28	Female	S. Korean	21	84	Music	2
ID09	29	Female	S. Arabian	18	-	-	3.5
ID13	19	Male	Brazilian	3	15	Agricultural Sciences	2
ID16	19	Female	Chinese	24	84	Accounting	2
ID21	20	Female	Brazilian	3	39	Biomedicine	2
ID22	23	Female	Brazilian	3	39	Biomedicine	3
ID24	21	Male	Brazilian	3	39	Animal Science	2
ID25	19	Female	Brazilian	3	36	Electrical Engineering	3
ID42	19	Female	Brazilian	3	60	Forestry	3
ID43	22	Female	Brazilian	3	24	Civil Engineering	3
ID45	23	Female	Brazilian	3	24	Agricultural Sciences	3
ID49	22	Male	Chinese	1	120	Finance	2
ID50	24	Male	Chinese	1	120	Urban Planning	2
ID51	20	Female	Chinese	1	96	Actuarial Science	2
ID59	21	Female	Brazilian	1	48	Biology	2
ID60	21	Male	Brazilian	1	72	Computer Engineering	2
ID73	20	Male	Chinese	24	24	Business Management	2
ID74	20	Female	Chinese	24	144	Accounting	2

Below I present the findings for the fifth research question, which derive from the post-writing questionnaire and from the semi-structured interviews. The questions in the post-writing questionnaire included multiple-choice questions and open-ended questions. The questions asked the participants which exam they thought was easier and which exam they liked more, both of which were followed up by open-ended questions asking the participants to explain why. Other questions asked what they thought was easy and what they thought was difficult about both exams, what they thought about the article, videos, and discussion, and so on (see Appendix D for all of the questions in the post-writing questionnaire). The questions in the semi-structured interviews asked about the students' overall opinions of the two exams, what they thought about the time limit, planning time, discussion, and source materials in the PBTW exam and what they thought about the time limit of the TW exam, and so on (see Appendix E for the list of questions asked in the semi-structured interviews).

4.2 RQ5: What are students' perceptions of the two different types of writing exams?

There were two sets of data that answered the fifth research question. The first data set originated from the post-writing questionnaire, which contained both multiple-choice and open-ended questions. The answers to the multiple-choice questions are quantitative and are presented first. The answers to the open-ended questions are qualitative and are presented second. For the participants' written responses to the open-ended questions, I only corrected spelling mistakes, but left other grammar mistakes as they were in the original responses. The second set of data which was used to answer the fifth research question is the semi-structured interviews. This data is solely qualitative and the results are presented third.

4.2.1 Post-writing questionnaire

Table 30 shows the responses that the participants made for the multiple-choice questions in the post-writing questionnaire.

Table 30
Answers to multiple-choice questions

Questions	**Answers**
Q1: Which exam was easier?	PBTW exam = 58 (72%) TW exam = 11 (13%) Both = 12 (15%)
Q3: What did you think of the videos?	Easy = 31 (38%) Difficult = 25 (31%) They helped think of ideas for the essay = 63 (78%) They did not help think of ideas for the essay = 12 (15%)
Q4: What did you think of the article?	Easy = 46 (57%) Difficult = 9 (11%) It helped think of ideas for the essay = 66 (81%) It did not help think of ideas for the essay = 7 (9%)
Q5: What did you think of the group discussion?	It helped think of ideas for the essay = 66 (81%) It did not help think of ideas for the essay = 10 (12%) It was not related to what I wrote = 5 (6%)
Q6: Did you use the ideas from the videos?	Yes = 48 (59%) No = 33 (41%)
Q7: Did you use the ideas from the article?	Yes = 67 (83%) No = 13 (16%)
Q8: Did you use the ideas from the group discussion?	Yes = 56 (69%) No = 24 (30%)
Q12: Which exam did you prefer taking?	PBTW exam = 62 (77%) TW exam = 18 (22%)

The first question in the post-writing questionnaire asked the participants which exam they thought was easier. Fifty-eight participants, or 72% of them, answered that the PBTW exam was easier, 11 participants, or 14% of the participants, answered that the TW exam was easier, and 12 participants, or 15% of the participants, answered that both exams were equally easy or difficult.

This question was followed by an open-ended question that asked why the participants thought one exam was easier than the other. Four main themes emerged from the participants' responses to the second question, which asked the participants to explain why they thought that one exam was easier than the other: ideas from sources and discussion, time, background information, and planning. Ideas from sources and discussion was mentioned by 33 different participants. Twenty-nine of the participants mentioned that they thought that the PBTW exam was easier because of the ideas from the videos, article, and discussion. Nevertheless, four participants had negative comments about the videos, article, and discussion. The participants' answers related to time mostly explained, with two exceptions, that they thought that the PBTW exam was easier because they had more time to take this exam. However, one participant wrote "Though there were lots of ideas and time seemed short" (ID01) and another wrote "in the shorter timed writing exam I could manage my time better" (ID04). The comments about background information describe that the videos, article, and group discussion in the PBTW exam gave the participants background information that they needed to understand the topic. Seven participants wrote that they thought that the PBTW exam was easier because they had time to plan their essay. However, one participant, the same one that mentioned he or she could manage his or her time better in the TW exam, wrote that he could plan his essay in the TW exam in much detail. He wrote "when I write the outline I can write very detail as main points, subpoints and how long the paragraph should have" (ID04). Table 31 lists the five themes that emerged from the second question, which asked why the participants thought one exam was easier, how many participants mentioned each theme, and sample comments related to each theme.

Table 31
Q2: Which exam was easier and why?

Themes	Times	Sample comments
Ideas from sources and discussion (positive comments)	29	ID05 "by watching the videos and reading the article, I got more ideas to organize my thoughts into my essay" ID23 "I could use something from the article and videos to build my essay" ID37 "The ideas from the discussion can provide and better developed essay" ID62 "after a discussion I will have more ideas on writing" ID80 "I can use the content in videos, lectures and discussion"
Ideas from sources and discussion (negative comments)	4	ID04 "in the longer one, I need to mix the idea of video and article that when I only wrote about 75% of my list (outline), the time already up" ID32 "I need write something from video or lectures" ID38 "I can not remember all the information from the videos and lectures" ID72 "I do not want watch the videos. Speak speed too quickly"
Time	23	ID19 "I could think more about my ideas and make sure what I wanted to write" ID 43 "we had more time to think about the topics" ID64 "I also have more time work on my essay" ID74 "we have more time to understand and think about our essay question" ID78 "More time to think about the essay"
Background information	22	ID27 "someone can write an essay, even if she/he does not have any background" ID29 "It is easier to write about what I have some previous knowledge" ID58 "the discussion and lectures could activate my background knowledge" ID75 "we understood the topic more clearly" ID84 "I get more information, background, and ideas from those materials"
Planning	7	ID 12 "The long timed essay is easier because it is easier to have better arguments and to plan the essay" ID15 "I had time to think carefully before start to write" ID26 "The longer timed writing was easier because I had more time to think about the subject and plan my essay" ID35 "it helped me to make plan" ID39 "Because I had more time to think about it and brain storm it"

The third, fourth, and fifth questions asked the participants what they thought about the videos, the articles, and the discussion in which they participated. Thirty-one participants thought

that the videos were easy, while 25 participants thought that they were difficult. Sixty-three participants answered that the videos helped them to think of ideas that they could use in their writing, whereas only 12 participants answered that the videos did not help them to think of ideas that they could use in their essay. Forty-six participants circled the answer that said the article was easy and 9 circled the answer that said the article was difficult. Sixty-six students thought that the article helped them to think of ideas, while only 7 participants thought that it did not help them to think of ideas. Sixty-six participants answered that the group discussion helped them to think of ideas for their essays and only 10 participants thought the opposite. Five participants answered that the discussion was not related to what they wanted to write.

Questions 6, 7, and 8 asked the participants whether they used the videos, article, and the ideas in the group discussion in their essay. Forty-eight participants answered that they used the videos in their writing, while 33 said that they did not do so. One participant did not answer this question. Sixty-seven students said that they used the article in their essay and 13 participants said that they did not. Two participants did not answer this question. Finally, 56 students said that they used the ideas they heard during the group discussion in their writing, but 24 of them said they did not. Again, two participants did not answer this question. Question 9 asked the participants why they chose not to use the videos, article, or discussion in their essays. There were three main themes for this question: opposing ideas, difficult videos or article, and time. The participants who mentioned opposing ideas all wrote that the ideas in the videos, article, or discussion were opposite to their own ideas and that is why they did not use them. Some participants wrote about the fact that they thought the videos and/or article was too difficult to understand and that is why they chose not to use them. Finally, some participants described that they did not have time to include or think about including the information from the videos,

article, and discussion. Table 32 contains each theme, how many participants mentioned them in their answers, and sample answers.

Table 32
Q9: Why did you not use the ideas from the materials or discussion?

Themes	**Times**	**Sample comments**
Opposing ideas	12	ID04 "I think the group discussion is useless when group members have different point" ID13 "the article and group discussion was not helpful because my point of view is different" ID28 "article only insists disagreeing opinion, so I didn't use it and group discussion too" ID40 "the ideas showed in those medias contradict how I think" ID58 "my point of view about the questions was different than the information in the videos, article and group discussions"
Difficult videos or article	9	ID06 "I did not understand the main idea of the video" ID07 "article vocabulary is difficult to me" ID25 "I did not understand very well the content of the video" ID71 "too fast to write down and hard to understand what they said" ID85 "we are not understand what the videos, article and ideas talked about"
Time	4	ID04 "I did not have enough time to write anything about the article" ID41 "I had not time enough to think about it while writing" ID53 "no time to write. There is too much resource" ID82 "I just wanted finishing my essay as soon as I can"

Question 10 asked the students to write about what was easy and/or difficult about the TW exam. There were six themes that emerged in the participants' answers to this question: background knowledge, time, planning, topic, arguments, and ideas. When the participants talked about background knowledge, they wrote that they did not have background knowledge on the topic about which they were writing, with one exception. One participant wrote "the knowledge I had it was enough to write a good essay" (ID44). The students who discussed time were divided: fifteen wrote that they did not have much time to write, while ten wrote that time was not an issue in the TW exam. All of the students who talked about planning said that they

did not have enough time to plan their essays. However, one participant said that she thought that it was easier to plan while taking the TW exam. She wrote "In shorter timed writing exam is easy for students get ideas and write about easy" (ID38). The next most commonly discussed theme was topic. Nine participants wrote that they thought the topic was difficult or they did not like it and five participants wrote that the topic of the TW exam was easier. Some participants also mentioned that they did not have enough arguments to write a good essay and others said that they liked that they could use their own ideas in the essay as opposed to having to incorporate sources. Table 33 shows the themes, the amount of times they were mentioned by the participants, and sample comments.

Table 33
Q10: What was difficult/easy about the TW exam?

Themes	**Times**	**Sample comments**
Background knowledge	27	ID01 "it was a topic I never thought about" ID03 "I did not know much about the topic" ID28 "if I have enough background of the topic, shorter timed writing exam would be more easy" ID39 "if is a topic I don't have much background my essay will be horrible" ID63 "I didn't know much about the topic"
Time	25	ID07 "time goes fast" ID13 "we do not have time to develop good/strong arguments in only 40 min" ID36 "sometimes I think I have no time to write whole time writing" ID41 "the time! It runs quickly, and I don't have time enough to formulate a good sentence" ID45 "write the text was too short" ID04 "I still can manage my time" ID79 "time was not a problem"
Planning	25	ID05 "the time was really short to brainstorm more idea" ID11 "was difficult to organize the ideas in short time" ID15 "it was difficult not have time enough to think about the topic before write" ID20 "you don't have enough time to plan and to think what you will write" ID47 "I could not organize the ideas"

Table 33 (Cont'd)

Topic	14	ID01 "topic was difficult" ID06 "the topic of shorter timed writing was difficult" ID54 "I didn't like the topic" ID12 "the subject was easier" ID31 "the subject was easy" ID33 "the prompt made it easier for me"
Arguments	11	ID03 "I could not write detailed arguments" ID12 "it is hard to (...) think in structure and not canned ideas to support the main point ID23 "the difficult was that I didn't have anything to support ID29 "it was hard develop relevant ideas in order to convince the reader" ID45 "I did not have data to support what I was writing"
Ideas	5	ID01 "the fact that I have develop the essay based on my own ideas was easy" ID08 "easy: Just write down what you think" ID18 "I already had an idea in mind after to read the prompt, which make it easy" ID32 "the easy things was I can use my own idea"

The next question asked the participants what was difficult or easy about the PBTW exam. The most commonly mentioned themes for question 11 were the following: incorporating sources, planning, topic, videos, and time. The most discussed theme in the participants' responses was incorporating sources. Half of the comments said that the participants did not know how to incorporate the sources in their writing and the other half wrote that the sources helped them to write better essays. Ten participants wrote that one easy thing about the PBTW exam was that they had time to plan their essay before they had the 45 minutes to write it. Eight students wrote about the topic of the PBTW exam: five students wrote that they liked the topic or that the topic was easy and three wrote that they did not like the topic of the topic was difficult. Seven participants said that they had problems understanding the videos. Finally, six participants had negative comments about the amount of time that they had to take the PBTW exam. Table

34 contains the five themes, how many times they were mentioned, and sample comments from the participants.

Table 34
Q11: What was difficult/easy about the PBTW exam?

Themes	**Times**	**Sample comments**
Incorporating sources	51	ID01 "to use videos and the article to support the idea was difficult" ID18 "too many information in the same time made me confused about what exactly I would write" ID31 "it was difficult to organize the ideas because the videos, article and discussion showed too much information" ID32 "the difficult thing was it had to choose which one is good for my essay ID71 "I didn't know how to cite the videos and article" ID03 "the use of sources such as video and the article helped me with the arguments" ID10 "I have a lot of information about the topic" ID15 "I had more support to think about the topic, with the article" ID26 "have all this information helped me to think faster and choose the information in easier way ID 47 "this variety of ideas helped me to support the main point"
Planning	10	ID11 "we have time to make an outline and think about" ID14 "I had time to plan and organize my ideas. It was so helpful and easier" ID20 "you have time to think and write calmly" ID34 "we had our time to think and organize our thoughts than we started our writing" ID46 "I could outline the points I would write"
Topic	8	ID01 "the topic was easy to develop" ID06 "it a topic that I knew" ID16 "The topic is harder than the shorter time easy" ID20 "I think the thesis was harder because gun control has to much controversy and I got in conflict with my ideas" ID26 "the topic did not catch my attention"
Videos	7	ID06 "the first video was difficult to understand its main idea" ID13 "the problem was the videos" ID23 "some part of the videos were a little difficult" ID28 "One of videos, agreement of gun use, was hard to understand to me" ID36 "they speak very fast and use some difficult words I can't understand

Table 34 (Cont'd)

Time	4	ID01 "it was difficult to manage the time" ID02 "it makes me think a lot until I lost my time" ID10 "I did not have enough time to mentioned all of them [the pieces of information from the sources] when I am writing" ID44 "The difficult was the time because we were under pressure of the time"

The twelfth question asked the participants which exam they preferred taking: the PBTW exam or the TW exam. Sixty-two participants, or 76% of the participants, answered that they preferred taking the PBTW exam. Eighteen participants, or 22% of the participants, said that they preferred taking the TW exam. Two participants did not answer this question. The thirteenth and last question asked the participants to explain why they preferred taking one exam when compared to the other. There were four main themes that emerged from the participants' responses for this question: ideas from sources and discussion, planning, background information, and test preparation. The participants who wrote about ideas from sources and discussion all said that the sources and group discussion helped them to think of ideas and arguments that they could include in their essay. Sixteen students wrote that they preferred the PBTW exam because they had time to plan their essay. Fourteen participants talked about the fact that the videos and article gave them background information to help them to write better essays. Some participants also mentioned that the TW exam helped them to prepare for other exams that they have to take, like the TOEFL (www.toefl.org) and the TW exam that the students have to take in ESL 220 and 221 at the ELC. Table 35 shows the themes, how many times the participants wrote about them, and sample comments.

Table 35
Q13: Which exam did you prefer and why?

Themes	**Times**	**Sample comments**
Ideas from sources and discussion	30	ID03 "in the second exam I had more material to base on. This way I could give more effective arguments" ID06 "the discussion and the lecture helped in good ideas and have more support in my ideas" ID22 "the longer timed writing permit open more points of view about subject and choose what the best way to argument and convince" ID49 "I got a lot of ideas from others" ID70 "the supporting material give me extra ideas and clues about what I am going to write"
Planning	16	ID14 "I can plan and organize my ideas" ID39 " give you (...) time to think about the topic and organize your ideas" ID43 "it is easier to develop my idea and organize it" ID64 "enough time to prepare" ID75 "we can make a draft first which make us have enough time to think and write down what we thought"
Background information	14	ID05 "I will be able to have background about the topic because some topics I might never hear or read about them" ID20 "you can practice all your skills to understand the topic" ID33 "it gave me more information about the topic, so I was more aware about the issue before starting to write the essay" ID44 "when we have the chance to know about it before we write the essay it is better for our performance" ID85 "this one [the PBTW exam] has more resources for me to understand
Test preparation	5	ID26 "it's better to prepare to future exams" ID67 "because for practice" ID80 "I can practice with both two exam" ID82 "because it is similar to TOEFL"

In short, 72% of the participants thought that the PBTW was easier and 77% of the participants preferred taking the PBTW exam. The answers to the open-ended questions in the post-writing questionnaire revealed that the students thought that the PBTW exam was easier and they preferred taking the PBTW exam for two main reasons: the articles, video, and discussion helped them think of ideas to include in their essays and the source materials gave the test takers

background information that they needed to complete the writing task. Fifty-nine percent of the participants reported using ideas from the videos in their essays; 69% of them reported using the ideas that they heard in the group discussion in their writing; and 83% of the participants answered that they used ideas from the article in their essays. Many participants wrote that they had difficulty integrating sources in the PBTW exam, partly because they could not understand the videos or article, or because they did not know how to cite the sources. On the other hand, a large number of the students wrote that the TW exam was difficult because they were not familiar with the topic or because they did not have enough time to write or plan the essay. The semi-structured interviews revealed very similar findings, but along with the post-writing questionnaire they add more depth to understand students' perceptions of the exams.

4.2.2 Interviews

Many of the issues discussed in the post-writing questionnaire also arose in the semi-structured interviews, such as difficulty incorporating sources in the PBTW exam, difficulty understanding the video, topic preference, time constraints, and planning time.

a) Difficulty incorporating sources

Three of the eighteen students who participated in the interviews mentioned difficulty incorporating sources in the PBTW exam. Below are their comments about this theme.

Excerpt 1

ID07 How can combine the video, article, and my idea

R So you think that is difficult to do?

ID07 Yeah, the combine. How can combine the video and article. How can supporting, how can example in my essay. I think is more difficult than the shorter timed writing.

(...)

ID16 I like the idea of discuss and talk about the idea, but not to support uh, like, take uh=

ID07 =how to support, we don't know

ID16 From the video, to build my idea, to have the concept of what I want to write

R So you would not like to have to use the ideas in your essay? It would be easier if you didn't have to use the ideas.

ID16 Yeah, it help me to think

ID13 Yeah, I think it should be optional. If you want to make your argument more strong you can use the videos or the articles. But if you don't want you can just write what you think.

Excerpt 2

ID60 Na segunda redação eu achei que eu recebi muita informação e não consegui organizar (...) Eu tinha muita informação mais os vídeos, os textos de apoio que eu li e eu meio que me perdi no tempo.

In the second essay [PBTW exam], I thought that I got too much information and I couldn't organize it (...) I had too much information plus the videos and the article that I read and I kind of lost myself.

In sum, three participants mentioned that they thought that combining the ideas from the article and videos with their own writing was difficult. Integrating the article and videos was not the only source of difficulty in the PBTW exam. Some participants also described having difficulty understanding the two short videos that they watched.

b) Using ideas from the source materials

Although the participants thought that incorporating the source materials in their writing was difficult, some participants mentioned that the videos, article, and/or discussion helped them generate ideas that they could use in their essay.

Excerpt 23

ID16 I think the second one [the PBTW exam] help me to make a contrast or idea (...) When we discuss about uh the topic, that will help me to think about what I want to write and my idea.

Excerpt 24

R What did you think of the two exams and which one did you prefer?

ID24 The long term.

R Why?

ID24 Because is easy to take a thesis statement when you talk to another students and read the article, because I used some parts of the article to make my essay.

ID21 For me even the long- the longer essay the topic was more difficult, for me was more easy because uh with the videos and the articles I had more ideas to write in my essay. So even the topic was more difficult for me was more easy.

(...)

ID22 I prefer the first one, because I could build my essay with this kind of argument in the article.

Excerpt 25

ID45 Deu mais suporte, a segunda porque teve os artigos pra ver- os vídeos, então acho que foi bem mais fácil que a primeira (...) As discussão com os grupos acho que também ajudou bastante porque deu mais idéias, pode trocar idéia.

It gave us more support, the second one [the PBTW exam] because there were the articles to see- the videos, so I thought it was much easier than the first (...) The discussions with the group also helped a lot, I thought, because they gave us more ideas, we could exchange ideas.

ID42 Aí vc percebe (incomprehensible) e muda o que você tava pensando. É bom pra ter mais exemplo, alguma coisa assim.

Then you realize (incomprehensible) and you change what you were thinking. It [the PBTW exam] is good to give us an example, something like that.

(...)

ID43 Eu acho que deu um exemplo assim no que basear pra escrever.

I think that it [the PBTW exam] gave examples to use as a basis to write.

ID42 Pra mim foi como um complemento, eu acho que não mudou totalmente a minha idéia, mas serviu pra complementar as idéias e tornar mais consistente, talvez, assim, tendo alguma coisa como exemplo ou alguma coisa que possa complementar seu pensamento pra deixar ele mais forte.

For me they [the videos and article] served as a complement, I don't think they changed my idea completely, but they served to complement my ideas and make them more consistent, maybe, like, having something like an example or something that can complement your opinion to make it stronger.

Excerpt 26

ID73 I would like to choose the first one, the long one, because we can watch the video and watch the article and talk about some information with the classmate so I can I think everyone can get more information and then to write a essay (...) I can use some (incomprehensible) from the article or the video to support my opinion in the essay.

ID49 The first exam [the TW exam] maybe you need to think about yourself and the second [the PBTW exam] you can discuss with other people so you can get ideas from other people and maybe it will give you new ideas and some support, something can support your idea.

All of the participants who mentioned using the source materials said that the article and videos gave them ideas that they used in their writing. In addition, some participants also had positive things to say about the group discussions. These students described that they used ideas that arose in the discussions in their essays.

c) Difficulty understanding the videos

Eight of the eighteen participants who were interviewed mentioned that they had difficulty understanding the video. Their opinions about the videos are expressed below.

Excerpt 4

R What about the videos, were they difficult?

ID07 The videos were so fast. The first guy is so fast so I just picking some ideas in my timed writing.

(...)

ID16 The video, the problem in the second exam, something in the video I couldn't understand.

(...)

ID21 I didn't understand a lot of the parts of the videos, but the article was very useful for me (...) Sometimes I didn't understand some parts of the videos, but I used what I understand. I thought it was a little difficult, the videos.

Excerpt 5

ID51 The second one [the PBTW exam], I think the problem is understanding, because we should understand the video and lectures and it's a little hard (...) I got the main ideas of the videos, but the details was a little hard.

(...)

ID49 The video maybe a little bit difficult because people speak very quickly and maybe use some word we didn't know.

(...)

ID50 I can't understand the video, only the main, not every part, every sentence.

Excerpt 6

ID74 The first one [the PBTW exam] we can watch the video, even maybe I don't really understand what's talking about in the video (...) The first video I didn't really understand what is he talking about, but I understood he agree with gun control. Or disagree, maybe? He thinks people should have gun, I already know this, but other points is hard for me to follow. And then the second video I understand most of the part.

(...)

ID73 For the video, uhm, I can, uh, get the main point, but I am not sure I can clear to get the detail information. Uhm, but at the essay [the article] is very easy. I can know the detail information.

The eight participants who reported having difficulty understanding the videos explained that the people in the videos spoke too quickly and that understanding the details of the videos was challenging, although they could understand the main ideas. One issue that emerged in the post-writing questionnaire was that of topic familiarity. In addition, the majority of the participants in the semi-structured interviews also discussed which topic they liked best.

d) Topic preference

Another theme that emerged in the semi-structured interviews was that of topic preference. Fourteen of the eighteen participants who were interviewed mentioned this theme. Four commented that they preferred the topic of gun control, while the remaining ten stated that they favored the topic of obesity. Below are the comments that the participants made about the two topics of the exams.

Excerpt 7

ID07 The topic is gun control, so is difficult to me (...) Change the topic [of the PBTW exam]. Topic is so heavy.

(...)

ID13 But I didn't like the subject [of the PBTW exam], that was gun control, because we spent three of four hours talking about this in the Speaking and Listening class so we already have a knowledge about this.

(...)

ID09 The topic [obesity] also very easy for me (...) It [gun control] is a big question for us, uh, because, uh, when I know this topic is about gun control I feel very scared. It's very scared.

Excerpt 8

ID24 I thought that the short timed writing exam was easier for me because the topic was easy and the long timed exam was easy too, but I felt more about the issue because it's not pretty easy to think about gun control and write in 45 minutes. I think that the problem wasn't a exam, but the topic.

(...)

ID21 For me, even the longer essay was, the topic [gun control] was more difficult, for me [the PBTW exam] was more easy.

(...)

ID22 The second one, the topic was easier than the first one (...) because gun control is a little difficult to express.

Excerpt 9

ID45 No Brasil a gente não discute muito sobre gun control, agora obesidade é mais discutido. Por isso eu acho que o segundo foi mais fácil.

In Brazil we don't discuss much about gun control, but obesity is much more discussed. That is why I think the second topic [obesity] was easier.

ID42 Pra mim também, porque inclusive a gente sempre ouve falar de obesidade daqui dos Estados Unidos, então é um assunto mais comum, então eu acho que é mais fácil.

For me that is true also, because we actually always hear about obesity in the United States, so the topic is more common, so I thought it was easier.

ID43 Eu também acho que obesidade foi mais fácil.

I also thought that obesity was easier.

Excerpt 10

ID59 Na primeira, eu tive muito mais dificuldade porque eu não tinha o que falar. Eu não, não era um tema que eu tinha a opinião formada nem em português. Tipo, se me mandasse escrever esse tema em português eu não saberia, eu ia ficar perdida também. Meus argumentos não tinha base, tipo, não tinha o que falar. Ficou ruim, tipo, eu sabia que eu tava escrevendo, mas tipo, eu não conseguia fazer melhor que aquilo porque eu não tinha o que falar sobre aquilo. A segunda não, a segunda, eu acho que, lógico, ter um preparo antes é muito melhor. Você, tipo, lê sobre aquilo, você tem mais embasamento pra falar. A segunda, acho que mesmo que se eu não tivesse, acho que teria ficado muito melhor que a primeira.

In the first one [the TW exam], I had much more difficulty because I didn't have anything to say. I didn't, it wasn't a topic I had an opinion about, not even in Portuguese. Like, if you told me to write about this in Portuguese I wouldn't be able to, I would be lost too. My arguments had no basis, like, I didn't have anything to say. The essay was bad, like, I knew what I was writing, but like, I couldn't do better than that because I didn't have anything to say about it. The second one [the PBTW exam], though, I think that, obviously having preparation before is much better. You, like, read about it, you have more support to use. The second, I think that even if I didn't have it [the support] I would have still written a better essay than the first.

(...)

ID60 No meu caso, eu me lembro que no ensino médio eu fiz uma redação parecida com a do desarmamento, no Brasil, aí eu tinha mais idéias e eu consegui colocar no meio tempo.

In my case, I remember writing a similar essay about gun control in high school, in Brazil, so I had more ideas and I could use them in the time I had.

Excerpt 11

ID74 The gun control one, I need to think about it more because I don't really familiar with gun control before I took the timed writing. But the one about junk food, I think that question is for everybody, cause it's like a social question.

(...)

ID73 From the topic, I think that gun control I can write more information than the junk food because gun control is a heat discussion in the United States and I studied gun control in the Speaking class also, so I can write more information in gun control.

Excerpt 12

ID51 I like junk food more. This topic, I have a lot of things to write, because is a (incomprehensible) always this topic and about gun control is also easier to write, but it's harder than the second one.

R So you thought gun control was more difficult?

ID51 Yeah.

ID49 Maybe I don't have any preference about the two topics (...) Junk food one may be easier one.

R Right.

ID50 I think the both topics is both okay, because if you explore this topic you will find that both topics are not just simple topic.

Only two of the eighteen interviewees said that they preferred gun control. Ten participants said they did not like the topic of gun control because it is a difficult topic or because they are not familiar with the topic. Nine participants mentioned that they liked the topic about obesity.

e) Time constraints

Five participants mentioned the issue of time. Most of them believed that they did not have enough time to write their essays during the TW exam. Below are the comments they made about the issue of time.

Excerpt 13

ID13 I think the short exam it's terrible because it's only 40 minutes and you don't have time enough to develop ideas good and teachers who collect our exam uh they want they require a high levels of our arguments and we don't have time enough to do a good strong argument in 40 minutes and it's really hard (...) I think the only problem [with the TW exam] is with the time. I mean, even in my native language, in 40 minutes I can't do a good essay, a strong essay with good arguments. And we are learning a new language. It's really difficult to put a good idea only in 40 minutes in both tests. I think more time would be a necessary choice.

Excerpt 14

R What did you think about the time limit?

ID24 It's [45 minutes] a good time to write, but a bad time to revise your essay. You can write in 45 minutes, but you can't revise your grammar mistakes, write and revise in 45 minutes. Probably my essay will be stuffed with a lot of grammar mistakes.

ID25 Yeah, we don't have time. I just write and revise never.

ID13 It's like crazy writing.

ID25 And I take long time to make the introduction because after I can't come back and change my ideas so I take a long time to start so at the final of the essay I don't have time to revise. That's it.

Excerpt 15

ID42 Eu sempre tive problema com o tempo, então fazer uma redação em 45 minutos pra mim é quase impossível porque meu cérebro pára e eu não consigo pensar em nada.

I always had problems with time, so writing an essay in 45 minutes to me is almost impossible because my brain stops and I can't think of anything.

(...)

ID 45 Quarenta e cinco minutos não dá pra você mostrar tudo que você sabe fazer, você só joga a idéia no papel, então acho que se tivesse mais tempo e se tivesse essas discussões, tipo, discutir o tema, acho que seria bem melhor.

Forty five minutes is not enough to show what you know, you only throw your ideas on the paper, so I think that if we had more time and if we had more discussions, like, discussing the topic, I think it would be a lot better.

(...)

ID42 Eu acho que o maior problema seria o tempo também (...) Eu acho que com o tempo você sente muita pressão e você acaba não escrevendo o que você sabe.

I also think that the biggest problem would be the time (...) I think that with the time limit you feel a lot of pressure and you end up not writing what you know.

Excerpt 16

ID51 I think the time is a problem because we have to uh think about the topic and writing in fifty minutes, fifty-five minutes (...) I write very slowly.

ID50 So do I, same question, the time problem. In the first uh exam [the TW exam] the information given in the article is limited so you should spend a lot of time to organize your thought. I think at last I have many thought I want to write them all down, but the time is limited. Uhm so if I have more uh time or more information uh maybe my writing is better.

All of the participants who mentioned the issue of time in the semi-structured interviews believed that 45 minutes was not enough time to write a well-developed essay. No one disagreed.

f) Planning time

Another theme that emerged in the semi-structured interviews was planning time. While some participants complained that they had no time to plan in the TW exam, others mentioned that the ten minutes given to them in the PBTW was also not enough time to plan. The excerpts below show the participants' thoughts about planning time.

Excerpt 17

ID16 My problem I want to write uh I always spend time to think what I want to write and make idea and think about uh the question (...) The first exam [the TW exam] it take a long time for me to think about what I want to write to find my idea.

Excerpt 18

ID22 Because [in the TW exam] we didn't have time to make a outline before we start the essay and I prefer always do first the outline.

Excerpt 19

ID43 A segunda teve mais tempo pra gente pensar o que a gente ía escrever, então achei mais fácil. A primeira deu o tópico e a gente já tinha que fazer o brainstorming, aí ficou mais difícil.

The second [exam] [PBTW exam] gave us more time to think about what we were going to write, so I think it was easy. In the first [exam] [TW exam], you gave us the topic and we had to start brainstorming right away, so it was more difficult.

Excerpt 20

ID59 A segunda eu acho que se eu tivesse mais tempo eu teria escrito mais, eu poderia ter pensado melhor em como eu teria dividido aquele texto, mas, tipo, nos 45 minutos eu fiz o que deu pra fazer.

I think that if I had more time in the second [exam] [PBTW exam] I would have written more, I would have thought more about how to organize the text, but, like, I did the best I could in 45 minutes.

(...)

R O que vocês acharam do tempo que vocês tiveram pra planejar?

What did you think of the time that you had to plan the essay [in the PBTW exam]?

ID60 Os 10 minutos?

The 10 minutes?

R É.

Yeah.

ID60 Eu acho que foi pouco pra mim. Eu tinha muita idéia e não consegui formar tudo.

I think it was too little for me. I had a lot of ideas and I could not organize everything.

ID59 Pra mim também foi pouco.

For me it was too little as well.

Excerpt 21

R What did you think about the planning time, the time you had to plan the essay?

ID49 I think the planning is very important. I always hard to study but when I have to write, when I have planning, I can write quickly. The first topic, gun control, we have less time to plan, so I think it's a little difficult to write the article, but the second one we have 10 minutes to planning. Uhm actually I don't use the uh this time, but I uhm, maybe just 5 minutes is enough because we know a lot of information from this topic.

ID50 Yeah, I think that the time before you writing is very important because you can write some key sentences, key words, to support your idea. Uh and like you have three key sentence, like, for example, so you know how to arrange your time reasonable. Maybe one point you will spend ten minutes, three sentence totally, so you can arrange your time.

Excerpt 22

ID74 Even I already understand, like, junk food is really a problem for world wide people, but I don't know how should I start cause I don't have enough time to write my guideline [outline] or anything for writing.

(...)

R What did you think about the time you had to plan?

ID74 Before we write our essay?

R Yeah.

ID74 Uhm, maybe we need some time to think about, cause the gun control you give us time to write not information, but our guideline in the paper, so when I start writing the paper I

know in the introduction what I want to write and the body or conclusion, what do I need to write. But when I write the one about junk food I can't, I think my essay is so messy. I think I don't have enough time to prepare.

R Yeah. Okay. Good.

ID73 Between the gun control and junk food writing, I think gun control I have lots of time to plan and to write and the junk food just 45 minutes, I only use 45 minutes to plan how to write and what should be write in the essay, so I think the gun control the time is better than the junk food. It's longer than the junk food.

Most of the participants who commented about the ten minutes that they were given to plan their essays agreed that they liked the time that they had to plan. However, two students thought that ten minutes was not enough to plan their essays.

Below is a summary of the main themes that emerged in the post-writing questionnaire and the semi-structured interviews:

1) The students like the planning time in the PBTW exam;
2) they believe that the article, videos, and group discussion are useful and give them ideas and background information that they can use in their writing;
3) some of them had problems integrating the ideas from the article and videos in their writing;
4) Some students found the videos difficult to understand;
5) most students liked the topic about obesity more than the topic about gun control;
6) they think that they cannot write a good essay in 45 minutes;

These are the test takers' perceptions of the two exams. Now I discuss what the raters think about the PBTW and TW exams.

4.3 RQ6: What are the raters' perceptions of the two different exams?

The two raters participated in two training and norming sessions, both of which were audio recorded. The first session lasted approximately two hours and the second one lasted one hour and thirteen minutes. The two sessions occurred three weeks apart. After the raters scored all 182 essays, I interviewed them individually, approximately three weeks after the second norming session. First, I report on the common themes that emerged in the training and norming sessions, which were rubric, source integration, and differences between the TW and PBTW essays. I then go on to report the themes that emerged in their interviews: the topics of the exams, source integration, the rubric, and the content validity of the exams.

4.3.1 Norming sessions

In the first session, both RK and RM shared their thoughts about the rubric, source integration and the differences between the TW and PBTW essays.

a) Rubric

RM did not like that the rubric assigned the same amount of points to each category, because she believed that content and organization should be more valued than spelling and punctuation. She said:

> "So each section is weighted evenly? Like, gets the same weight, so like, they don't get extra points doing a good job on content (...) Most writing rubrics I've seen give more weight to, like, content and organization and lastly to spelling punctuation."

RM mentioned this same issue in the second session as well. She said:

> "And and I just, like, the writing teacher in me thinks it's just not fair that punctuation gets the same weight as like content. I just think that that's a problem."

RK mentioned the fact that the rubric only had three levels, although there were actually four levels and one could assign a zero, a one, a two, or a three. She explained that, since no one would receive a zero, the rubric really only had three scores that could be assigned to one's essay.

> "That's another problem with the rubric, you know, only having- I guess there's four options, but essentially three. I mean, no one's going to get a zero. It's like 'oh, is it a low two or a high one?'"

RM agreed with RK, and said the following in the very beginning of the second norming session:

> "So one thing is this is a tough rubric because it doesn't weigh things and there's not enough spread so I feel like, particularly, like, there's a lot of essays where you're like 'okay, well, this isn't a three and it's not a one so it has to be a two, but there's a lot of, like, probably a lot of spread and equality of twos across the board (...) There's not enough room in each category to say, like, like you basically have good, bad or in between and there needs to be more spread in the in between part."

In the middle of the session, she again brought up the lack of spread in the rubric:

> "And they're interrelated and this is where spread would help cause if there were- if there was more spread in the middle, you could say well this is- for instance, this one is much

better organized than the one we just read right before it because it, like, has an introduction, it has body paragraphs, so you could give credit for having topic sentences, but no thesis if there was more spread in the middle."

RK agreed with RM and suggested that the rubric might be better if it had half scores, for instance, if one could assign a score of 2.5.

In the second session, RK also mentioned that she had difficulty with the wording of the rubric. She said:

"And I had trouble with the wording, like some or frequent, like, we decided on kind of numbers for spelling, but grammar, what's frequent?"

Both RK and RM described that they had difficulty assigning scores for cohesion. RM said:

"Cohesion is hard to- we're not sure that, like- partly the rubric is not very helpful with cohesion, like, satisfactory cohesion, and we talked about that that's a sentence level issue, but in some of these essays it's hard to tell whether it's a problem with cohesion, whether it's a problem with organization or globally, whether it's a problem with grammar that, like-"

RK interrupted RM and added the following:

"I found myself giving relatively- well now that I'm looking at it, maybe they're not- kind of relatively high cohesion scores just because even if you have a lot of grammar mistakes, if I can figure out what you're saying in each sentence most people can, like, link them together in a logical way."

RM also said that assigning scores for vocabulary was not easy, particularly deciding between a two and a three. She said:

"This is the hardest category for me to know the difference between a two and a three because, like, so, like, he- for instance, he misuses principle in here, which I would argue for a two, but on the other hand, broadcasted. He doesn't usually seem to have problems with vocabulary for what he's trying to say, but then he has, like, you know, word class issues, like unhealth."

RK added:

"I think it's hard because there's so much to look at for this one category: repetition, and do they use it right."

RK revealed that organization was also a difficult category for her to assign a grade. She said:

"So this is another category on the rubric there's just so much to look at, you know, there's so many factors."

Both raters had very negative comments to say about the Weir's (1990) rubric. They thought that the rubric should have a wider range of scores and clearer descriptors for each score and category.

b) Source integration

I asked the raters what they thought about the PBTW exam, which they had just finished scoring. RM discussed that the participants could not incorporate sources very well. She said:

"Yeah, no one can do that [incorporate sources] very well (...) They just don't know how to introduce it and they don't know what to do with it once it's there so, like, and this is a problem- so this is a problem with the essays more generally is like, a lot of people just jump in 'in the video' and you're like 'wait, what video?' I mean, I know what video they're talking about, but more generally your audience doesn't know that, but that's-

that's kind of a global problem anyway, like, ESL writers have a hard time with context. Actually, native speakers have a hard time with context. My undergrads can't do that either, but somehow with the sources it's more glaring when they're talking about sources because there's no context for who the people are."

RK responded:

"They all seem to grab the same stuff."

Both raters agreed that the test takers lack ability to integrate sources in their writing. RM felt very strongly about this issue.

c) Differences between the TW and PBTW exams

At the end of the second norming session, I asked the raters if they noticed any differences between the TW essays and the PBTW essays. RM answered:

"So I kind of hate to say it, I think in some ways these [the TW essays] are better because there aren't as many- like, the other ones if you could incorporate sources well it helped, but there were so many cases in which they didn't do it very well, that, like, it was it was worse, whereas this is like just purely opinion-based. It takes out the, like, 'they don't actually know how to use sources.'"

Both RK and RM thought that the TW essays that they read in the norming session were longer than the PBTW essays. RK and RM's exchange about the length of the essays is below:

RK These are longer, I think.

RM These do seem to be longer.

RK Maybe it's cause they're just kind of getting all their opinions out.

RM explained that she preferred reading the TW essays mainly because it did not require the students to incorporate sources and the students' lack of ability to do so was not so distracting. Both RK and RM thought that the TW essays were longer than the PBTW essays.

4.3.2 Interviews

The raters were interviewed separately one day after they had finished scoring the essays. I first interviewed RK on Skype and then RM in person. I asked the raters the same questions and below I report on the common themes that emerged in both interviews, two of which were also mentioned in the norming sessions: the topics of the exams, source integration, the rubric, and the content validity of the exams.

a) Topics of the exams

One of the questions asked the raters if they believed that the two topics were comparable. Both of them seemed to think that they were indeed comparable. Nevertheless, they thought that gun control was a bit more complex than obesity. Below is what RK had to say about the topics:

> "Yeah, because they're both kind of like a controversial topic, gun control more so, though. I feel l like whenever you bring up gun control it's just such a loaded topic, you know. Compared to the obesity thing, where they can just kind of say what they think you want to hear or whatever. People have stronger opinions about guns."

RM answered the question in a very similar manner, but she elaborated more on why she thought obesity was an easier topic:

"I think, overall, I guess they were comparable. I do think, though, that the obesity topic was easier for people to just talk from their own experience than gun control, so you tended to, I don't know if this is a good or bad thing, but you tended to get more, like, kind of opinion statements, or experience-based statements in the obesity stuff, in the obesity papers, than the gun control paper. So for instance, there were a handful of gun control papers that talked about Chinese kids saying their Chinese families had- were concerned about living in the U.S. because of guns, but that was like two or three in the whole lot, whereas the obesity papers tend- you tend to have a lot more like, you know, 'I was surprised when I came to the U.S.' or 'American stores do this,' and you don't necessarily need- you just need experience to talk about that, you don't need much more knowledge so to speak."

The raters agreed that the topic of gun control was more complex for international students than the topic of obesity.

b) Source integration

Both raters mentioned the difficulty that the participants in this study seemed to have integrating sources in their writing. RK said the following:

"When they're asked to use sources it might be too much in, like, a timed environment, because they're already supposed to make this organized essay and have good details and support a good argument and then you also have to throw in this other aspect [source integration]."

RM said:

"I do think that for the PBTW exams people who were able to integrate information probably did better in terms of content because it gave them something to talk about, but most people weren't able to integrate information, so then it ended up being like, it didn't end up helping, I guess, because you get, like, a random fact, and no discussion of it."

While both raters thought that the participants in this study could not integrate sources well, RM believed that the few participants who could integrate sources performed better in the PBTW exam.

c) Rubric

Both raters stated that they had problems with the rubric, an issue that first arose in the second norming session, which occurred three weeks before the interviews. Once again, RK and RM shared that they thought that the descriptors in the rubric did not distinguish themselves well from one another. RK said:

"There's kind of a big jump between some of the scores, the description. It was hard for me, like, vocabulary was hard to decide. What's 'frequent,' what's 'some,' what's 'almost no inadequacies?'"

RM's comments about the rubric's descriptors are below:

"I didn't like the rubric because I think the rubric didn't do a good enough job of like distinguishing—like, it needed to have more of a range because I think—I felt like I ended up with a lot of stuff in the two category because it wasn't good but it wasn't terrible, but things would be rated twos for different reasons (...) More precise descriptors would've helped distinguish, like, you know, orga- organization, for instance, there's lots of ways you can deal with organization, so just like having first, second, third doesn't

necessarily mean well-organized, but I tended to give, like, priority to that because at least a rough structure, but then, like, it doesn't mean that the ideas actually flow, right?"

RK also discussed which categories were easy and difficult to score with the analytic rubric. She thought that grammar, punctuation, and vocabulary were the most difficult, and spelling and organization were easier. This is what she said about the categories:

"Spelling was the easiest. So the categories that I found I put off when I was going through each essay, like the three categories that were the last ones for me to decide the score for, grammar, punctuation and vocabulary. That was just for me, I don't know. It took more time, you really had to look through the whole thing, and it didn't stand out as much. Something like organization or content kind of jumps out from the beginning, for me at least (...) And punctuation I thought was ha- kind of difficult because you really have to, like, look at every different sentence and, okay, maybe there's a few run-ons in the whole essay, does that mean it's- does that drop it down to a two? I feel like there was, oh, the content, it's kind of like, yes, they addressed the prompt, but it was just so surface level. Is that still a three? I don't know."

Once again the raters mentioned that the descriptors in the rubric should be clearer.

d) Content validity

RK and RM also mentioned the issue of content validity. Carr (2011) defined content validity as "how well the test sampled the subject matter or curriculum on which the test was based" (p. 152). In other words, if a test has content validity, it measures exactly what the teacher is teaching. The two raters believed that the PBTW exam was a good tool to evaluate students in an academic course in which source integration is one of the objectives. RK stated:

"If that is one of the main objectives of that class, to be able to do that [integrate sources], then I would say that it is an advantage to see whether they can. So for me, I guess, comes down to is that something that they've really been taught how to do beforehand, and if so, then I guess that would be a preferred format."

Below is what RM said about content validity:

"Yeah, so I, like, I actually like the idea of the PBTW exam because I think- I think it has the potential to measure kind of higher level synthesis writing skills in a way that the just regular one can't. So particularly if you are looking for people's ability to like write from sources, to handle kind of more complex academic writing, like, I think that that's actually a better- like, it's a better task because it more naturally mimics the type of stuff you have to do because you have to pull information from a variety of different sources."

The raters thought that the PBTW exam is a good tool to evaluate students when the course objectives include teaching them to integrate sources in their writing. A summary of the raters' opinions of the TW and PBTW exams gathered in the norming and training sessions and in the interviews were the following:

1) The rubric lacked range and clear descriptors;
2) the students did not know how to integrate sources in their writing;
3) the TW essays were longer than the PBTW essays;
4) gun control was a more complex topic than obesity;
5) and the PBTW exam is valid when the course objectives include source integration.

Now that I have presented the results of the quantitative and qualitative research questions, I jointly interpret and discuss the results of the quantitative and qualitative

data, following the convergent parallel design described by Creswell and Plano Clark (2011). The reason for discussing the results of the study after the quantitative and qualitative data have been presented is to allow me to gather different pieces of information about the TW and PBTW exams, such as scores, the lexical and syntactic complexity, as well as the grammatical accuracy and fluency of the essays, and the students' and raters' perceptions of the exams, all of which, combined, paint a more in-depth picture of the two exams in question.

CHAPTER 5: DISCUSSION

In this chapter, I discuss the results of the quantitative and qualitative research questions. I do this while following the mixed-methods design that Creswell and Plano Clark (2011) described as the convergent parallel design, in which the researcher combines the quantitative and qualitative results in the interpretation phase of the study. First, I briefly summarize the main findings of the quantitative and qualitative questions. Second, I argue why PBTW exams are the best way to evaluate ESL academic writing based on the results of my study. Third, I discuss the students' and (fourth) the raters' perceptions of the two exams. Fifth and finally, I finish this chapter by discussing the difficulties of implementing PBTW exams in ESL academic writing courses.

5.1 Main findings of the study

The foremost finding of this study is that PBTW exams have many benefits over TW exams. For example, when students are given time to familiarize themselves with the topic of a writing test through video-watching, readings, and discussions, and when they are also allowed time to plan, they write significantly longer essays and significantly more words per minute. These results were also found by Cumming et al. (2005), David (under review), Ellis and Yuan (2004), Kellogg (1988), and Ong and Zhang (2013). Longer essays are good because they have been found to be strongly correlated with higher scores, as reported by Guo, Crossley, and McNamara (2013). Guo et al. (2013) found that text length was a strong predictor of the quality of an essay for both independent and integrated writing tasks. In addition, in the PBTW exams, the participants used significantly more sophisticated vocabulary and more different word types

and types of nouns in the PBTW. This benefit was also found by Ong and Zhang (2013) and Cumming et al. (2005), who stated that learners used a wider variety of words in their study on the effects of task complexity. These findings actually contradicted work by Johnson et al. (2012), Kormos (2011), and Kuiken et al. (2005). Perhaps the exposure to the source materials and their classmates' ideas allowed the learners to use new vocabulary that they encountered in the reading, videos, and the discussions with their classmates. Indeed Cumming et al. (2005) hypothesized that the learners in their study used more different words in the integrated task because they borrowed words from the source text. Moreover, the planning time might have given students the opportunity to think more carefully about the vocabulary that they were going to use in their writing, which may have also been the case Ong and Zhang's (2013) study.

The participants in this study also scored significantly higher in content and punctuation in the PBTW exam. This could be explained by the higher amount of time the students spent on task in the PBTW exam and by the many ideas learned from the source materials and brainstormed in the discussion and planning time. Before the students wrote their essays in the PBTW exam, they spent 45 minutes learning and discussing about the topic. The high level of engagement with the task in the PBTW exam, combined with the extra planning time, probably prepared the test takers better for the writing task and allowed them to brainstorm more ideas to include in their writing. The extra planning time also gave test takers the opportunity to think about how to organize their essays more and focus solely on writing in the 45 minutes that they had to write. In contrast, the learners had to plan, organize, and write in 45 minutes in the TW exam. Last, but not least, using ideas from the source materials in the PBTW exam may have been the reason why they scored higher in content for this exam.

But the case for PBTW exams over TW exams is not 100% closed. In some ways, students performed better or the same in the TW exam. Results in this study were similar to other studies (Kuiken, Moss & 2005; Tedick, 1990; Winfield-Barnes & Felfeli, 1982) in which learners did not write more accurate essays when they had to write more involved essay-writing tasks. Their spelling scores were significantly higher in the TW exam. Furthermore, the participants in this study did not score significantly higher when they took the PBTW exam, findings which are similar to those of other studies (Plakans, 2008; Cumming et al., 2005). This could suggest that the discussion, reading, and videos aided the participants' in writing essays with higher quality *content*, but perhaps with less focus on spelling (less technical essay-writing accuracy). One interesting finding of this study was that the participants did not perform differently in terms of syntactic complexity, unlike the results that I found in my earlier study about process-based exams (David, forthcoming). Research on how learners' writing differs in terms of syntactic complexity when writing in different genres suggest that their writing is more syntactically complex when they write argumentative essays when compared to narratives (Lu, 2011; Polio & Yoon, 2014; Way et al., 2000). However, the results of this study suggest that this is not the case when learners are writing across the same genre, but about different topics and under different conditions (TW exam versus PBTW exam). It seems as though what helps learners write more syntactically complex essays is not the condition in which they write, but in which genre they are writing. Many impromptu timed-writing exams, as I suggested above, elicit essays about personal experiences to ensure that the participants are not given a prompt with which they are not familiar. Writing about personal experiences might generate essays that are similar to narratives and, as research suggests, learners seem to use simpler syntax when writing narratives.

Although there were no significant differences between the scores that the participants received on both exams, some participants scored much higher in the PBTW exam while others scored considerably higher in the TW exam. Therefore I decided to select the participants who scored much higher in one exam to investigate why that was the case. Participants ID12 (Brazilian), ID39 (Brazilian), and ID52 (Chinese) all scored relatively higher in the PBTW exam. They scored 5.5, 4.5, and 4.5 points more in the PBTW exam, consecutively. After comparing the scores that they received for each category, I observed that all three of them scored 1 point or higher for content. ID39 and ID52 received scores twice as high for content in the PBTW exam. These two participants also scored 1 point or higher for cohesion in the PBTW exam. ID12 scored twice as much on punctuation and 1 point more on vocabulary in the PBTW exam and ID 39 scored twice as much for organization. All of the other scores that the participants received were either the same or half a point higher. These three participants received much higher scores for content in the PBTW exam, findings that the results that the *t* tests also indicate.

Participants ID31 (Brazilian), ID44 (Angolan), and ID 72 (Chinese) scored 4.5, 5.5, and 7.5 points higher in the TW exam, consecutively. All three participants scored 1 or 1.5 points more for vocabulary and 1, 1.5, and 2 points more for spelling. Indeed the results of the *t* tests revealed that the participants scored significantly higher for spelling in the TW exam. However, it is surprising that these three participants received higher scores for vocabulary, because the results of the *t* tests revealed a different picture: the participants used significantly more types of nouns and words, as well as more sophisticated words in the PBTW exam. Perhaps the raters' perceptions of "better" vocabulary are different from the criteria that RANGE (Nation, 2005) and the Lexical Syntactic Analyzer (Lu, 2012) use to measure lexical complexity. Two participants, ID44 and ID 72, scored 1 point and 1.5 points higher for cohesion and ID31 and ID72 scored 1

point or 1.5 points higher for content. A trend seems to emerge from the analyses of these six participants: most of them received higher scores for content and cohesion. These two constructs are perhaps linked to higher quality essays.

These six participants' preferences for one exam seem to align with their performance in the exams. Two of the three participants who scored considerably higher in the PBTW exam answered that they preferred the PBTW exam when compared to the TW exam. The participant who preferred the TW exam, however, answered that the PBTW exam was easier. Similarly, two of the three participants who scored much higher in the TW exam said that they preferred the TW exam. This might be an indication that the participants' preferences could influence their performance. In fact, 73% of the participants who scored higher in the PBTW exam reported that they preferred the PBTW exam.

The students in this particular study may have performed better on spelling on the TW exam because they had more resources for paying attention to these mechanical forms. As Robinson (2001) explained, a task can become more complex when learners are required to complete other tasks related to the main task. During the PBTW exam, participants in this study had to discuss, watch videos, and read an article before they began writing. All of these tasks may have made the PBTW more complex than the TW exam. The complexity may have made it slightly different in the eyes of the test takers. The need to integrate sources and use source information may have made the test takers think less about focusing on spelling. Rather than focusing on the minute mechanics of writing, in the PBTW exam, test takers may have focused on getting their ideas across and using source materials more. In taking time to plan and formulate their ideas, they may have concomitantly focused more on content and punctuation (as the data suggests) because these may correspond with structuring (well) an essay. Baralt,

Gilabert, and Robinson (2014) explained the following about task complexity and unfamiliarity with task types:

> "Repeated attempts to perform complex tasks will prompt the use of more complex language in such a way that the proposed effects of task complexity on 'pushing' the complexity of responses to task demands, and 'streching interlanguage,' are more obvious" (p. 19).

Perhaps the students' unfamiliarity with the procedures of the PBTW exam may have taken a toll on their performance and if they have more opportunities to practice this type of process-based task more often they will use more complex language.

There is more evidence that the PBTW exam and the TW exam do not tap into the exact same construct. The scores that the participants received in the two exams correlated only moderately, demonstrating that they have low concurrent validity (they do not measure the same thing). This is a finding also revealed in Cumming et al.'s study (2005) and David's (under review) study. Moreover, the correlation coefficients for the categories of vocabulary, punctuation, and spelling were also moderate, while the coefficients for content, organization, cohesion, and grammar were small. The impromptu TW exam may only measure writing (or may measure it more narrowly), while the PBTW exam likely measures a more broad construct of writing, with the following constructs additionally tapped into: reading, listening, and source integration.

However, this integration of tasks may be new to some of the test takers. Being a good reader (being able to comprehend a text) and being a good writer does not necessarily translate into being able to use information from a reading passage while writing. The integration of different tasks is difficult and must be practiced if one wants to be good at it. Such findings were

revealed in Delaney's (2008) study about reading-to-write tasks; good readers still have to learn to write what they have read about. Thus, not only are the two test types potentially measuring differing constructs, but one (the TW exam) may be familiar to the students in this study, while the other (PBTW) may be unfamiliar. This further complicates a clear view of the differences between the two exams.

Another finding of this study is that the overall intra and inter-rater reliability for the PBTW exam was slightly lower than the reliability for the TW exam. When analyzing the inter-rater reliability for each of the analytic categories, I found that grammar and punctuation and vocabulary and punctuation had the lowest reliability in the TW exam and PBTW exam, respectively. On the other hand, spelling and cohesion and spelling and content had the highest reliability in the TW and PBTW exam, respectively. The reason for these findings may be related to the rubric used in this study. The rubric and its effect on reliability and raters' perceptions of the rubric will be discussed later. But another reason again may be task familiarity. The raters in this study were more used to rating TW exams. Their unfamiliarity with scoring PBTW exams may have led to their lower agreements in scoring them. After all, raters get better with practice (Weigle, 2002).

Asking participants what they thought tended to also point to the superiority of the PBTW exam. The great majority of the participants in this study thought that the PBTW exam was easier (72% of the participants said so in the post-writing questionnaire), and they preferred the PBTW exam in comparison with the TW exam (77% of the participants said so in the PBTW exam). The reasons for this preference varied. The most mentioned reasons were: 1) the article and videos gave the participants many ideas that they could use in their writing (a theme that also emerged in the semi-structured interviews); 2) they had more time to engage with the writing

task; and 3) they were more familiar with the topic because of the article, videos, and discussions. Eight of the eighteen students interviewed thought that the videos and article helped them to think of ideas that they could use in their essay. However, some participants explained that the article, videos, and group discussion were not very helpful when their ideas opposed the ideas presented in the sources. In addition, some participants thought that the videos or article were difficult to understand. This theme also emerged in the semi-structured interviews. Eight of the eighteen participants who were interviewed said that the videos were very difficult to understand. Again, this comment tends to show the true integration of the task, because when asked about the PBTW exam, test takers talked about all aspects of the exam, including the videos.

When asked about the TW exam, the participants mainly complained about the fact that they were not familiar with the topic and did not have time to plan their writing, as well as the short amount of time that they had to write the essay. As reviewed at the beginning of this paper, researchers know that students perform better on TW exams if they are given time to plan (Ellis & Yuan, 2004; Kellogg, 1988; Worden, 2009). And researchers have described how students score higher when they are tasked with writing about something that is familiar to them, as opposed to something unfamiliar (He & Shi, 2012; Tedick, 1990; Winfield-Barnes & Felfeli, 1982). Thus the students' complaints are valid. Five of the eighteen students interviewed said that they did not have enough time to write the essay in the TW exam and ten explained that they either did not have enough time to plan their essay in the TW exam or in the PBTW exam, even though they were given 10 minutes for planning.

The raters also had a clear preference for the PBTW exam. They indicated that they thought that the PBTW exam had more content validity if one of the objectives of a course is to

teach students to integrate sources. At the same time, the raters complained that the majority of the participants did not know how to integrate sources, and one of the raters explained that the participants who could integrate sources did it extremely well and received a higher score for content. However, she felt that this was rare and in most cases the lack of ability to integrate sources was distracting. This may show again that the test takers are unfamiliar with the PBTW format. The raters' confusion on how to score non-integration of sources may also be one of the sources of their lower reliability with PBTW exams.

5.2 Why PBTW exams are a better fit to evaluate ESL academic writing

The construct of academic writing is not one that is easy to define. While there are many models that attempt to explain the writing process, not many attempt to define academic writing. However, all of the models of the writing process, including the Hayes-Flower (1980) model, the Hayes' (1996) model, and Bereiter and Scardamalia's (1987) structure of the knowledge-telling model, include elements such as background information, planning, drafting, and revising. The ACTFL can-do statements for advanced academic writing include some of these skills as well (see http://www.actfl.org/sites/default/files/pdfs/Can-Do_Statements_2015.pdf for the complete list of can-do statements for all proficiency levels), in addition to other skills. Below are some of the can-do statements for advanced learners (these range from low-advanced to high advanced):

- I can revise class or meeting notes that I have taken for distribution
- I can draft and revise an essay or composition as part of a school assignment;
- I can write a research paper on a topic related to my studies or area of specialization;
- I can write a position paper on an issue I have researched or related to my field of expertise;

It is clear from these can-do statements that skills such as planning, revision, and research are ones that advanced L2 writers should master. The CSU Expository Reading and Writing Task Force put at the top of their list of the writing skills that college-level students must have to succeed in regular academic classes the following two writing skills: synthesizing ideas from different sources and integrating quotations and citations (as cited in Ferris, 2009). Moreover, Weigle (2002) noted that academic writing in higher education is very often based on a reading or listening text and that most assignments that college students complete require them to incorporate sources, which can be the course textbook or readings the students have researched. Weigle's observation was more than a simple observation; it is also supported by research, as discussed above. Research on the types of writing tasks that students do in tertiary non-ESL academic classes often include research papers, critiques, summaries, and so on (Cooper & Bikowski, 2007; Hale et al., Horowitz, 1986; Yigitoglu, 2008). In addition, advanced university L2 writers have to know how to analyze, interpret, create, and summarize information, persuade others about their opinions, and conduct and write research projects (Grabe & Kaplan, 1996).When suggesting activities that advanced L2 writers should engage in, Grabe and Kaplan listed planning, using multiple sources, reading critically, engaging in guided discussion, writing outlines, and being exposed to multiple types of writing genres, none of which are promoted or encouraged by impromptu TW exams.

Cumming (2013) called for a redefinition of the writing construct for English for Academic Purposes classes that includes source integration. Most ESL writing textbooks, such as Sourcework: Academic Writing from Sources (Dollahite & Haun, 2007) (used at the English Language Center), teach students pre-writing strategies and source integration, and many writing teachers encourage and might even require their students to plan, write an outline, and do

research on a topic before they begin writing the first draft of a paper. With the exception of in-class exams, many professors assign papers that students have to write outside of class. These papers allow students with plenty of time to plan and do research.

A definition of the construct of academic writing, based on the information above, then, should include the following elements, among others: planning, drafting, revising, researching, synthesizing ideas, thinking critically, and integrating sources. TW exams clearly do not support the use of these writing skills. The results of this study suggest that the TW and PBTW exams possibly measure two different constructs, findings that Cumming et al.'s (2005) and David's (under review) studies also revealed. Indeed it is not difficult to conceptualize why integrated and process-based exams do not measure the same constructs as impromptu TW exams. The impromptu TW exam measures writing, while the integrated and process-based exams measure the following constructs in addition to writing: reading (the participants had to read and understand the article); listening (the participants had to watch two videos and understand the information in them); and source integration (the participants had to integrate information from the article and videos in their writing). These are all skills that are required of students in college-level classes, according to the definition of academic writing that I proposed above.

However, these were not the only skills the participants had to use in the PBTW exam. They were encouraged to take notes while watching the two videos and not every student is a good note taker. One participant wrote in the post-writing questionnaire, "[the videos were] too fast to write down and hard to understand what they said" (ID71), suggesting that he was not able to take notes while watching the videos. Some participants did not even take notes, although the teachers who helped to collect data and I always encouraged them to do so. One participant wrote, "I cannot remember all the information from the videos" (ID38), which probably indicates

that she did not take notes. In addition, the participants were given time to plan. If students are not accustomed to planning and not familiar with pre-writing strategies they may not be efficient planners. The participants in this study may not value pre-writing during writing exams because of the nature of the impromptu TW exam with which they are accustomed. Such exams put emphasis on the writing product, not the writing process. Worden (2009), for example, found that learners who engaged in high levels of pre-writing performed significantly better, suggesting that knowing how to apply pre-writing techniques to planning and engaging in higher levels of pre-writing does indeed affect the learners' final writing product.

As explained above, many models of the writing process and ESL writing textbooks include pre-writing as an important skill in the process of writing. While some students may engage in planning even in impromptu TW exams, most will most likely not do so because of time constraints. Indeed many students in this study complained that they did not have time to plan during the impromptu TW exam. ID20, for instance, said about the TW exam, "you don't have enough time to plan and to think what you will write." Some even complained that ten minutes of planning in the PBTW exam was not enough. Engaging in group discussion before writing can also be seen as a pre-writing activity because it gives students the opportunity to exchange ideas and think of new ideas.

One key component of the PBTW exam is source integration. The participants had to read an article, watch videos, and synthesize and integrate the information that they learned in the source materials in their writing. Most experts in the field of L1 and L2 writing agree that these skills are crucial for students, native or nonnative, to succeed in their non-ESL academic classes, as described above. More than half of the participants mentioned in the post-writing questionnaire that they had difficulty integrating the information from the sources in their essays.

This same issue arose in the semi-structured interviews. ID07, for example, said that she thought it was difficult to combine the videos and article with her ideas. The students were not the only ones who mentioned their difficulty integrating sources. The raters also noticed that the participants did not know how to integrate sources in their writing. RM mentioned that most of the participants could not successfully incorporate sources in the essays that she read. RK agreed and added that the participants seemed to all use the same information from the sources. Second language learners' difficulty integrating sources is an aspect of integrated tasks that other researchers have found (Cumming et al., 2005; Gebril & Plakans, 2013; Sawaki, Quinlan & Lee, 2013). Sawaki et al. (2013) wrote the following about source integration: "The task of integrating information from various sources into written discourse requires a complex coordination of language modalities" (p. 93). They suggested that test takers who cannot integrate sources successfully may do so for two reasons: they do not understand the source materials, which may have been the case in this study, or they may have problems choosing relevant information from the source materials and organizing them in their own writing, which may also have been the case for the participants in this study. While some might argue that L2 learners' difficulty incorporating sources is a valid argument for not including integrated or process-based writing tasks in progress or achievement exams, others might argue the exact opposite: If a program's objectives include teaching students to engage in the writing process by reading, discussing, and planning, and if the curriculum includes synthesis and source integration, then to obtain an accurate measure of students' progress and/or retention of course objectives, the impromptu timed-writing exam is not appropriate. Hargis (2003) noted that exams should "become the curriculum" (as cited in Carr, 2011, p. 55). Carr explained that if a program has a communicative curriculum, then the students should be assessed accordingly. If a program teaches writing as a

process, and teaches other skills related to academic writing, such as synthesis and source integration, then the students should be tested on these very same constructs, and the process-based writing exam provides teachers a better picture of whether students have mastered these skills.

As long as ESL programs continue to evaluate learners' writing by using impromptu TW exams, teachers will continue to train students to take such exams. This, in turn, creates a negative washback in L2 writing classrooms, which forces teachers to ignore (at least for part of the course) the construct of academic writing on which they should be focusing. Washback, as defined by Carr (2011), is how a test affects classroom teaching and learning. While teachers could be teaching students all of the writing skills listed above, such as planning, synthesis, source integration, critical thinking, and so on, they are wasting precious classroom time by training their students to take impromptu TW exams. Weigle (2004) investigated a new placement test that integrates reading and writing and found that the test has already created a positive washback in the classroom, with teachers now focusing more on critical thinking skills and text analysis. Carr (2011) warned that independent-skill tests "puts pressure on teachers to focus on developing discrete skills in isolation so as to better prepare students for their tests" (p. 17), but writing is not a skill separate from reading, listening, and speaking. Weigle (2002) added that, especially for classroom assessment, impromptu TW exams give students the false impression that the writing process is not really important, especially when students' main concern is to perform well on the test.

The fact that the students in this study were not comfortable integrating sources in their writing can be evidence of the negative washback in their academic writing classes. The students who participated in this study, in particular, have to take two timed writing exams per semester,

both of which combined are worth 20% of their final grade. Teachers and students alike want to ensure success in these exams and value in-class timed writing activities as a result. The College Conference on Composition and Communication stated the following about writing assessment: "Writing assessment is useful primarily as a means of improving teaching and learning" (as cited in Deane et al., 2008, p. 66). If teachers and language testers start implementing PBTW exams, academic writing teachers will start teaching these crucial writing skills that students so desperately need in order to succeed in their non-ESL academic courses and therefore improve their teaching.

Impromptu TW exams clearly do not mirror the skills that L2 learners will need to succeed in their ESL academic writing classes or in regular academic classes that they will take nor do they mirror the elements of academic writing that I discussed above, such as planning, synthesis, and source integration. While PBTW exams do not allow learners to revise, they give them the opportunity to engage much more in the writing process than impromptu TW exams, by allowing learners to learn and discuss about the topic and plan their writing. In addition, these exams require learners to read and think critically about the topic, which are also crucial skills for regular academic classes. Last, but not least, process-based writing exams allow learners to practice and demonstrate their source integration skills, skills which are much valued both in the ESL academic classroom and in other academic courses.

5.3 Students' preference

Very few studies have investigated test takers' opinions about writing exams (David, under review; He & Shi, 2008; Lee, 2006; Powers & Fowles, 1999) even though their perceptions are extremely important because, according to Rea-Dickins (1997), they are one of

the main stakeholders. In addition, if learners do not like an exam (if they feel the exam is inauthentic in any way; see Carr, 2011, p. 160), the learners may not feel invested in it, and their lack motivation to prepare for the exam and take the exam may affect their performance (Lee & Coniam, 2013). This issue of how test takers and teachers perceive tests is called *face validity* (Hughes, 2003), and face validity is related to the perceived authenticity of a test (Carr, 2011). That perception can be self-evolved or adopted through the expressed beliefs of others (other students, the teacher, or other stakeholders). Students may singly or collectively have a negative view of a test if they believe that it does not measure what they are learning or if they perceive that others do not value it or trust its scores. The information gathered in the post-writing questionnaire and in the semi-structured interviews suggests that the PBTW exam appears have more face validity than the TW exam, which aligns with results from David (forthcoming) and Lee (2006). The participants in this study had a preference for the PBTW exam for the following reasons: (a) they learned background information and heard new ideas, both of which they could use in their writing; (b) they had time to plan their writing; and (c) they had more time to engage with the topic of the writing exam.

One way in which the source materials helped learners was by giving them background information about the topic and providing them with ideas to use in their writing. Many participants said that they liked the PBTW exam better because it gave them background information through the reading and the videos. ID05, for instance, wrote the following in the post-writing questionnaire about this issue: “I will be able to have background about the topic because some topics I might never hear or read about them.” Furthermore, many participants agreed that the source materials helped them to think of ideas and arguments to use in their essays. This was evidenced by the fact that more than 75% of the participants answered that the

article and videos helped them to think of ideas for their writing. Many of them also brought up this issue in the open-ended questions and in the semi-structured interviews. ID03 wrote in the post-writing questionnaire, "in the second exam I had more material to base on. This way I could give more effective arguments." ID45 said in the semi-structured interview, "It gave us more support, the second one [the PBTW exam] because there were the articles to see- the videos, so I thought it was much easier than the first." Research has suggested that topic familiarity may affect test takers' writing performance (Ellis & Yuan, 2004; Kellogg, 1988; Worden, 2009). When students who are unfamiliar with a topic are given texts to read and/or videos to watch, they can learn the necessary background information that they need to perform well. These source materials can also help them to write essays with higher quality content, as the results of this study suggest, because of the ideas that they can borrow from the videos and reading.

Another important component of the PBTW exam was planning time. The participants had ten minutes to plan their essays. While some participants complained that ten minutes is not enough to plan an essay, the majority of the participants thought that allowing them to plan their essays before they began writing was a good idea. ID26 wrote in her response to an open-ended question in the post-writing questionnaire: "The longer timed writing was easier because I had more time to think about the subject and plan my essay." Some participants also mentioned this theme in the semi-structured interviews. ID45, for example, said, "I think the planning is very important. I always hard to study but when I have to write, when I have planning, I can write quickly. The first topic, gun control, we have less time to plan, so I think it's a little difficult to write the article, but the second one we have 10 minutes to planning." As mentioned before, providing learners with planning time may affect the overall quality of their writing, as many studies have found (Ellis & Yuan, 2004; Kellogg, 1988; Worden, 2009).

Group discussions are a huge part of communicative ESL classes and other non-ESL academic classes in the United States. Group work is so important for academic success that the Common Core Standards said the following about preparing students for college classes: "To build a foundation for college and career readiness, students must have ample opportunities to take part in a variety of rich, structured conversations—as part of a whole class, in small groups, and with a partner" (National Governors Association, 2010). Even though Shi (1998) did not find significant differences between the group who participated in pre-writing discussion and the group who did not, group discussions allow students to exchange ideas, to formulate their own ideas and opinions, and to be aware of opposing ideas, all of which can be beneficial to writing. More than 81% of the participants answered that the group discussion helped them to think of ideas to write in their essays. ID62 wrote in one answer to an open-ended question: "after a discussion I will have more ideas on writing." ID45 said the following about the lack of group discussion in the TW exam: "Forty five minutes is not enough to show what you know, you only throw your ideas on the paper, so I think that if we had more time and if we had more discussions, like, discussing the topic, I think it would be a lot better."

Similarly, in the study to investigate a process-based writing placement exam mentioned above, Lee (2006) discovered that many students found the ten-minute group discussion extremely useful. They mentioned things very similar to the participants in this study, such as the benefit of being exposed to new ideas and brainstorming new ideas to include in their writing. The same results were found in my study comparing PBTW and TW exams (David, under review) mentioned above. Seventy percent of the participants preferred the PBTW exam and six of the forty participants mentioned that they especially liked the group discussion because it gave them new ideas and it allowed them even more time to brainstorm ideas. The results of this

study, combined with the results of Lee's (2006) and David's (under review) studies suggest that students clearly believe that group discussions and planning time have a positive impact on their writing. Group discussions also mirror what happens in communicative ESL classes and other non-ESL academic classes in American colleges and universities. Program and test administrators might be concerned that if students discuss their ideas before they write, they might borrow ideas from their classmates and their writing might not reflect exactly the test takers' opinions. However, discussing ideas with other people is a normal social aspect of writing, as described by Prior (1998). Weigle (2002), for example, criticized the lack of a discussion component in the TOEFL's writing tasks because, according to her, this makes the test less authentic because discussing about a topic before writing is a normal part of most academic classes. Lee (2006) describes the writing process as "a process of discovery in which ideas are generated and not just transcribed as writers think through and organize their ideas before writing and revising their drafts" (p. 307). Exchanging ideas in groups is one of the ways learners can generate ideas.

Not all of the students' responses and comments about the PBTW exam were positive, however. Some participants complained that the videos were too difficult and that the people in the videos spoke too fast. Thirty-one percent of the participants answered that the videos were difficult in the post-writing questionnaire, and, when explaining why they did not use ideas from the videos or article, eight of the participants wrote about their difficulty understanding the videos. ID06, for instance, wrote, "I did not understand the main idea of the video" and ID71 wrote, "too fast to write down and hard to understand what they said." Seven participants also mentioned that the videos were hard to understand when they discussed what was difficult about the PBTW exam in the post-writing questionnaire. Moreover, eight of the eighteen participants

who were interviewed also brought up this issue. ID21 said, "I didn't understand a lot of parts of the video" and ID49 said "The video maybe a little bit difficult because people speak very quickly and maybe use some word we didn't know." These results are not surprising. In their study about the new integrated writing tasks for TOEFL's Test of Written English (www.toefl.org), Cumming et al. (2005) found that the test takers used the reading passage much more often than the video lectures. They hypothesize that this was due to the fact that the learners had access to the reading passage while they were writing, but they could not go back to the video lectures and they had to solely rely on their memory. Because reading is self-paced and learners can go back to sentences they cannot comprehend, reading may be an easier skill than listening. Some participants even mentioned that they did not use the ideas in the videos because they could not remember them or they had not taken notes. Another negative aspect of the PBTW exam that the participants mentioned was their difficulty integrating sources, as discussed above.

Most of the participants who were interviewed thought that obesity was a less complex topic than gun control and some of the participants' majors might have been a cause for that. Three of the four participants were from Science-related majors, such as Biology and Biomedicine and therefore might have been more familiar with a health-related topic such as obesity than gun control. However, even students from other majors, such as Accounting or Engineering also agreed that gun control was more complex. Perhaps the issue was more related to cultural backgrounds than majors. Controversial topics such as gun control are heavily influenced by culture. Gun control is a topic that is much discussed in the United States, but not so much in other countries where gun laws have been in place for a long time. In Saudi Arabia and Brazil, for example, guns are banned, so there is not much controversy on the subject.

Obesity is more of a worldwide problem and people might have more background information about it than gun control. What was interesting was that students who took the PBTW exam on gun control and the TW exam on obesity and the students who took the PBTW exam on obesity and the TW exam on gun control both agreed that gun control was more difficult, even after they received background information about the topic. This could be further evidence that the participants might have found gun control more complex because of their cultural background.

As mentioned above, Rea-Dickins (1997) believes that test takers' perceptions of exams are difficult to investigate and to put into practice. However, when the results are this clear, and when other studies support these same findings (David, under review; Lee, 2006), it does not seem too difficult to make use of their perceptions. If students complain that they do not have enough time to plan their essays in the TW exam or even in the PBTW exam, and if they believe that planning is an important step in the writing process, which is the case for the participants in this study, why not give them time to plan their writing in writing exams? Perhaps it might even be useful to include pre-writing techniques in the exam so that the test takers can choose the technique they will use. That way, the test takers will be exposed to different pre-writing techniques and be encouraged to use them. Another one of test takers' perceptions which is simple to address is the fact that they valued and made use of source materials. If the test designers or teachers have to account for time constraints, instead of providing students with one article and two videos, they can give students one shorter article or show one short video. They can even ask the test takers to read or watch the source materials at home, before the exam. When students take exams in non-ESL academic classes this is exactly what they do. They study for the exam by reading their textbooks and lecture notes. Group discussions are also not challenging to include in a writing exam. A short ten or five-minute group discussion could get

students' ideas started before they begin planning their writing. This could even be done one class before the exam. Finally, the participants in this study said that spending more time on task was beneficial to their writing process, another element of the PBTW exam which is not impossible to achieve in a writing exam if test designers and teachers include source materials, group discussions, and planning time in writing exams, all of which allow learners to engage more with the exam.

The students are not the only ones whose perceptions matter when it comes to writing assessments. Teachers' perceptions of the writing tests being used to place and evaluate students in their writing courses are also important (Cumming, Grant, Mulcahy-Ernt & Powers 2004). Although it was beyond the scope of this study to investigate teachers' perceptions of writing exams, Weigle (2002) noted that teachers are worried about test usefulness. They are concerned, for example, whether the test being used in their class accurately tells them if their students have reached the goals of the course, how the results of the test will help students to become better writers, and whether students are interested in the prompts created for their writing exams. If the goal in an academic writing class is to determine whether students see writing as a process, whether they can apply this process to writing, and whether they can integrate sources in their writing (all of which many writing experts agree are academic writing skills, as seen above), then the impromptu TW exam may not be telling the teacher whether the students have reached the goals of the class.

In this particular study, there were no significant differences in the students' performance in the TW and PBTW exam. However, their preference for the PBTW exam was evident based on their responses in the post-writing questionnaire and in the semi-structured interviews. There is no doubt that the participants value the writing process through which the PBTW exam allows

them to go. They may value this process because of its higher content validity. In other words, the process-based writing exam mirrors more closely what students do in their ESL academic writing classes. It may seem unfair to students that their teachers (including myself, as I was the teacher of more than thirty of the participants) encourage them to see writing as a process and engage in the writing process in each assignment that they must complete, but do allow them to go through this process during the two timed-writing exams that they have to take. Finally, the process that the students go through in the PBTW exam may be the reason why the participants in this study scored higher for content in the PBTW.

It is important to acknowledge that my role and identity as teacher to some of the participants and as compatriot to all of the Brazilian participants may have played a part in the way that they interacted with me and responded to my questions in the post-writing questionnaire, but especially in the semi-structured interviews. Richards (2003) warned qualitative researchers that identity can have an effect on how the interview unfolds. My ESL 220 students might have been more inclined to say what they thought I wanted to hear, as opposed to how they really felt, a practice commonly found in qualitative research. Indeed, as Richards (2003) explained, "we cannot ignore our relationship with the interviewee and the effect this might have on the way the talk develops" (p. 85). He stated that interviewees respond to questions by taking into account the interviewer's cues, even if those cues are not apparently noticeable. I did not share my feelings about the two exams with my students, but that does not mean they might not have been sensitive to them. In addition, the data were collected at the beginning of the semester and the students were going to continue classes with me for at least two or three more months. It is undeniable that I held a position of more power than them, which could have contributed to them saying what they thought I wanted to hear. As compatriots, the

Brazilian participants, especially the ones who were interviewed in Portuguese (and who were not my students), might have been more inclined to say how they felt because of our shared experiences as Brazilian students in the United States and because I was not their teacher. I was an outsider with shared cultural beliefs, experiences, and background. Indeed I found that one particular group of Brazilian participants who were interviewed together had more lengthy responses than most participants. Furthermore, even though we did not know each other before data collection, all of the Brazilian participants who were not my students asked me questions about my personal life, my life in the United States, my plans for the future, and so on, before and after the interviews. This might be evidence that they might have been more comfortable talking to me than other participants. Other participants from countries other than Brazil, including the ones who were my students, did not behave the same way. The Brazilians who were not my students might have seen me as somewhat equal to them because they knew that I was a student at the university and they were students as well. This may have allowed them to share more about their opinions.

Finally, the students in this study seemed much more interested about the prompt related to obesity than the one about gun control, and topic preference could influence students' motivations for writing, which in turn could potentially have an effect on students' performance (Weigle, 2002).

5.4 Rubric design and use

One important issue that I did not originally plan on investigating was rubric design and use. However, because the raters constantly mentioned the rubric during the norming and training sessions and during the interviews, it is difficult to ignore it. The raters who participated

in this study both complained that the rubric lacked range. Weir's (1990) rubric only has three different scores, although the scores can range from zero to three. Not one of the test takers in this study received a zero for any category, which means that the rubric really only ranges from one to three. Regarding this issue, RM said, "I ended up with a lot of stuff in the two category because it wasn't good but it wasn't terrible, but things would be rated twos for different reasons." RK said something very similar: "That's another problem with the rubric, you know, only having- I guess there's four options, but essentially three. I mean, no one's going to get a zero." Knoch's (2009) rubric created for a diagnostic test, for example, contains nine possible scores. Georgia State's Test of English Proficiency (GSTEP), as found in Weigle (2004), is also an example of a rubric that has multiple levels of scores. Raters can give scores that range from one to ten. TOEFL's rubric for the integrated writing task also has a wider range of scores, which can go from zero to five (see www.toefl.org for more information).

Another problem that the raters mentioned about the rubric was that it lacked clear descriptors for each category and each score level. RK, for example, said, "And I had trouble with the wording, like some or frequent, like, we decided on kind of numbers for spelling, but grammar, what's frequent?" and RM said, "More precise descriptors would've helped distinguish." Weir's (1990) rubric has the following descriptors for organization: 0) No apparent organization; 1) Very little organization of content. Underlying structure not sufficiently controlled; 2) Some organizational skills in evidence, but not adequately controlled; and 3) Overall shape and internal pattern clear. Organizational skills adequately controlled. Words such as "little" and "some" are very vague and do not give raters a clear picture of what is expected of the test takers. The rubric for the GSTEP includes more specific information about what it means for an essay to be organized, with phrases such as "introduction and conclusion present, but may

be brief" and "connections between and within paragraphs are made through effective and varied use of transitions and other cohesive devices" (Weigle, 2004, p. 50). The lack of range and clear descriptors in Weir's (1990) rubric might have influenced inter and intra-rater reliability.

It is difficult to ignore that RM seemed to play a much more dominant role during the training and norming sessions when compared to RK. RK was much quieter and many times simply agreed with RM without elaborating on her opinion very much. Perhaps RM's more dominant role may have led RK to share less about her opinions or to agree with RM and avoid confrontation. It was impossible, however, to have separate training and norming sessions because the objective of these is to increase rater reliability and agreement. RK did, however, share more of her opinions during the interview, which was conducted separately. Her perceptions of the rubric and the exam during the interview seemed to match what she said during the training and norming sessions.

The inter-rater reliability coefficient for both exams was somewhat high. Carr (2011) recommended an alpha level of .800 or higher as acceptable for high stakes exams and an alpha level of .700 or higher for low stakes exams. However, the coefficient for the TW exam was higher (.728) when compared to the one for the PBTW exam (.643), which is not acceptable for low stakes exams, according to Carr. Similarly, the coefficients for intra-rater reliability were higher for the TW exam. RM had coefficients of .352 for the PBTW exam and .862 for the TW exam and RK had coefficients of .642 for the PBTW exam and .552 for the TW exam. Weigle (2004), however, found the opposite. The integrated task in her study proved to generate more reliable scores than the independent task. Nevertheless, the researcher might have obtained these results because the raters could only assign two scores: pass or fail. The TW exam may have been simpler to score because the raters did not have to deal with source integration.

Although the rubric did not contain any categories for source integration, it seemed that the raters were attending to it nonetheless. Comments such as "I do think that for the PBTW exams people who were able to integrate information probably did better in terms of content because it gave them something to talk about, but most people weren't able to integrate information, so then it ended up being like, it didn't end up helping, I guess, because you get, like, a random fact, and no discussion of it" and "they all seem to grab the same stuff" point to the raters' attention to how the participants integrated sources in the PBTW exam. RM, especially, seemed distracted by the participants' inability to integrate sources, which could have affected the way she scored the essays and explain the lower inter and intra-rater reliability coefficients for the PBTW exam. The raters' attention to the test takers' lack of ability to integrate sources could have affected the way that they scored the PBTW essays. Perhaps the raters were trying to compensate for the test takers' lack of ability to integrate sources by assigning lower scores to other categories in the rubric, such as content or organization. RM herself mentioned that the students who could integrate sources successfully scored higher in content, which could be evidence that the ones who did not received lower scores for content even though the descriptors in the rubric did not say anything about source integration. In the training session, the two raters and I discussed how the raters should attend to source integration and we agreed that test takers who could not integrate sources or who chose not to use any of the source materials would not be punished in any way for doing so. However, as the raters' comments suggest, this did not seem to be the case.

It was quite surprising that the inter-rater reliability coefficient for grammar for the TW exam was extremely low (.012). Other studies have found low reliability coefficients for grammar as well. Winke (2013) found reliability coefficients of .49 for grammar in the context

of group oral exams. She argued that this coefficient is low and that it is not an effect of the test, but it is a problem that the raters had when using the rubric. The raters in her study reported having difficulty focusing on grammar while also having to focus on fluency, vocabulary, and overall communication skills. They also reported not thinking that grammar was of particular importance to the tasks that the learners performed. The researcher suggested that eliminating grammar from the rubric might be the solution, because, according to her, it might "free up their mental resources for discerning the other categories" (p. 262). Although her recommendations are for oral exams, this may be a good idea for writing exams as well, especially because professors who teach non-ESL classes might not be concerned about L2 learners' grammar mistakes, but about their ability to demonstrate content knowledge. Furthermore, the interlanguage of a learner may develop slowly. It can take time for learners to improve grammatical accuracy and many learners may even plateau. As Gass and Selinker (2008) explained, sometimes even if learners are frequently exposed to the L2, their interlanguages may still plateau, which further complicates the development of interlanguage grammars and the discussion the importance to evaluate it in academic tasks.

Grammar was not the only low reliability coefficient in this study. The intra-rater coefficient for organization for the TW exam for RM was .000 and the coefficient for spelling for RK for the PBTW exam was .167. It is difficult to understand the reason for the low spelling coefficient, especially because the raters and I established limits for spelling mistakes that fit each of the three possible scores, as described above. We decided that if the participants had less than five spelling mistakes, they would receive a 3; if they made more than five mistakes, they would receive a two; and if they made a spelling mistake in every other sentence, they would receive a one. RK herself seemed to think that rating spelling was not difficult, because she

mentioned in the interview, "Spelling was the easiest." RM mentioned in one of the norming sessions that she had difficulty distinguishing organization from cohesion, so we had a discussion with RK and decided that organization was global and cohesion was local. However, it seems like even after we agreed on a more specific definition of the two categories, RM still had difficulty assigning a score for organization. Both raters have had experience scoring essays, but RK has had more experience scoring listening tasks. When scoring such tasks, the rater does not have to pay attention to spelling and this may be one reason why the intra-rater reliability coefficient for spelling was so low.

Perhaps the main reason why the intra-rater reliability coefficients were not very satisfactory was because we should have spent more time training and norming. We trained and normed for over three hours, but that may not have been sufficient time for the raters to become familiar with the rubric. Studies about rater reliability, such as Weigle's (1998) and Congdon and McQueen's (2000) studies, had training sessions that lasted approximately half a day, which was considerably longer than the two training and norming sessions in this study combined. The other issue was that there were only two training and norming sessions, but the raters took approximately six weeks to score all 182 essays, which could have affected their consistency rating the essays. Indeed Lunz and Stahl (1990), Lumley and McNamara (1995), and Congdon and McQueen (2000) found that raters' judgments can fluctuate from one rating session to another. Weigle (2002) suggested that raters be given a set of essays for calibration at the beginning of each rating session when rating occurs over the course of more than one day. Perhaps with a longer training session, more regular norming sessions, and daily essays for calibration, the intra-rater reliability might have been higher.

5.5 Hurdles of implementing PBTW exams

Implementing PBTW exams in ESL programs is not an easy task, and that is why many programs still use the impromptu TW exam to assess learners' language abilities and monitor their achievements. The first obstacle in the implementation of PBTW exams is in designing the test itself. To design a PBTW exam, one must not only select a topic and write a prompt, but he or she must also find readings and/or videos that give test takers background information about the topic and ideas that they can use in their writing. In addition, the readings and videos have to be level and age-appropriate and short enough to be read in a limited time frame. If the test is to be similar to the one used in this study, with a brief discussion at the beginning to activate learners' knowledge of the topic, the test designer has to create questions. If one wishes to administer two exams per semester, as the English Language Center does, this entire process would have to be done twice. To make things even more complicated, the test designer would probably have to create multiple, equated exam forms so that he or she can have different forms of the exams to choose from for each administration and to ensure that students do not know which form of the exam and which prompt they will answer. This is a way to avoid plagiarism and the memorization of entire essays, a practice that He and Shi (2008) found was common among the 16 Chinese students that they interviewed in their study about TOEFL's Test of Written English (see www.toefl.org for more information on the test). In addition to all of these steps, the test designer has to go through all sorts of recommended procedures for creating an assessment tool, such as writing test specifications, pre-testing prompts, choosing a rubric, and so on (Weigle, 2002).

Designing the test is not the only problem in the implementation of the PBTW exam. The other hurdle is scoring the exam and interpreting students' scores, as Carr (2011) explained. He

noted that, with tests that integrate reading and writing, when a student does well, that probably means that he or she is a good reader and a good writer. However, what happens when he or she does poorly? Does he or she have poor writing skills, reading skills, or both? In the case of this particular PBTW exam, which requires the participants to know how to use pre-writing techniques and how to integrate sources, to have good listening skills and note-taking skills, the issue is even more complex than Carr described.

Cumming (2013) responded to Carr's concern by saying that the only problem with "task dependency," that is, when students' skills in one task (reading, for example), can have an effect on their performance on the main task (writing), is when one assumes that writing is an independent skill. The results of Esmaeili's (2002) study on reading-to-write tasks suggest that reading and writing are extremely interrelated. In fact, it is impossible to see any of the four language skills as an isolated skill. When learners do a listening activity in class, they may take notes to remember the details or main points in the listening passage, combining listening and writing into one task. When learners are speaking to their classmates or teacher, they also have to use their listening skills to understand them and respond appropriately. When students are writing a paper, they may read their paper twice or more to revise and edit, or they may do research and read about the subject of their paper, using their reading skills in addition to their writing skills. In their regular, non-ESL, academic classes, students read book chapters and articles and then take an exam in which they most likely have to answer questions in writing about the content of the textbooks. Not to mention the fact that most writing assignments that professors require their students to do include reading and writing skills, such as writing a summary, a critique, or a research paper. In short, as Cumming (2013) explained, students'

ability to integrate information in their writing is exactly the skill that ESL academic writing courses need to measure.

One further difficulty with implementing the PBTW exam is ensuring that the students' proficiency level is high enough that they will actually be able to read and listen to the source materials and understand them, which according to Cumming (2013), is crucial for the students to succeed in process-based writing tasks. The fact that the students in this particular study did not perform differently in the TW and PBTW exams may reflect this challenge. Perhaps the students in this study have not reached a proficiency level that allows them to successfully read and listen to source materials and successfully integrate them in their writing. In fact, some students brought up this exact issue in the post-writing questionnaire and semi-structured interviews. Thirty-one percent of the participants answered that the videos were difficult in the post-writing questionnaire and others mentioned this same issue in the semi-structured interviews. As I already mentioned, Cumming et al. (2005) found that test takers use reading passages more often than listening passages because they could refer back to the article, but not the video. Perhaps test designers should choose reading passages over listening passages when creating integrated tasks.

Almost 61% of the participants mentioned that they had difficulty integrating sources in their writing, and three of the eighteen participants who were interviewed also had something to say about their difficulty with source integration. Furthermore, both raters seemed to voice their concerns about this issue in their interviews, as mentioned before. RM, in particular, described students' difficulties integrating sources twice. She said, "but most people weren't able to integrate information" in one occasion and "no one can do that [incorporate sources] very well (...) They just don't know how to introduce it" in another occasion. RM even mentioned that

native speakers of English have problems incorporating sources. The students who participated in this study had all been taking 220 or 221 for less than a month at the time of data collection. One reason why they might not have done well integrating sources in their writing could be because the teacher had not covered source integration in their classes yet. Had the data been collected at the end of the semester, the results may have been different.

Ensuring that all test forms are equivalent in terms of difficulty level is a clear challenge for creating process-based and integrated writing tasks, as Weigle (2004) noted. She warned test developers that it is difficult to always choose reading passages of the same linguistic difficulty. Perhaps one way to avoid choosing reading passages of different proficiency levels is to compare the passages to a corpus to make sure that the passages are similar in terms of the number of words and word families that they contain. RANGE (Nation, 2005) would be a good tool for that purpose. Another issue with designing integrated tasks that Weigle (2004) mentioned is choosing topics that are equivalent. A solution to this, according to the author, is to write detailed test specifications that can be easily followed.

There are many hurdles to implementing PBTW exams. However, the students' preference for these exams and the fact that they have more content validity when it comes to the goals of academic writing classes alone are good reasons to implement such exams for progress and achievement purposes. Regarding content validity, both raters agreed that PBTW exams are a valuable tool to evaluate writing when course objectives include source integration. RM, for instance, said, "I actually like the idea of the PBTW exam because I think- I think it has the potential to measure kind of higher level synthesis writing skills in a way that the just regular one can't. So particularly if you are looking for people's ability to like write from sources, to handle kind of more complex academic writing, like, I think that that's actually a better- like, it's a

better task because it more naturally mimics the type of stuff you have to do because you have to pull information from a variety of different sources" and RK said, "If that is one of the main objectives of that class, to be able to do that [integrate sources], then I would say that it is an advantage to see whether they can. So for me, I guess, comes down to is that something that they've really been taught how to do beforehand, and if so, then I guess that would be a preferred format." For them, content validity should be taken into consideration when designing a test, and many test designers would agree (McNamara & Roever, 2006; Weigle, 2002). If an ESL academic writing course teaches students that writing is a process that includes reading, discussing, planning, and so on, it is only fair that the exams used to check students' progress during that course assess these same skills. During the semi-structured interviews with the eighteen test takers, I asked which exam they thought was more similar to the things that their instructors were doing in their 220 or 221 classes. All of the participants agreed that the PBTW exam reflected more what they did in class than the TW exam. Other reasons include the fact that the participants in this study and other studies used more sophisticated vocabulary and wrote longer essays in the PBTW exam, not to mention that they scored higher for content.

CHAPTER 6: CONCLUSION

The findings of this study revealed many interesting aspects of the PBTW exam and its apparent advantages over the impromptu TW exam, especially in the case of achievement testing after academic writing classes have been taken by the students. The participants in this study wrote significantly longer essays and used significantly more sophisticated words in the PBTW exam. They also used significantly more word types and significantly more different types of nouns in the PBTW exam. The high level of engagement with the exam and the addition of source materials may have contributed to these results. The participants most likely borrowed ideas and words from the reading passage and the videos, which could have resulted in longer and with more sophisticated and varied lexicon. In addition, they received higher scores for content in the process-based exam, which again could be explained by the source materials. The videos and article may have contributed with background information and ideas that the test takers could use in their writing, which may have affected the superior content of their essays.

Another finding of this study was that the participants thought that the PBTW exam was easier and they clearly displayed their preference for this type of exam over the impromptu TW exam. Reasons for this preference included the fact that they could use the ideas in the source materials and discussion as background information and supporting points for their arguments, as well as the extra time that they were given for planning. The raters also displayed their preference for the process-based writing exam when the goal of the course is to teach students to integrate sources. The process-based exam has more content validity to evaluate the construct of academic writing described above, which includes planning, research, source integration, among other elements. However, both the participants and the raters mentioned the difficulty that the

test takers had integrating sources in their essays. Source integration is a complex skill that required practice, but at the same time it is an important skill to measure in academic writing classes. The raters expressed their dislike for the rubric. They believed that the rubric lacked range and specific descriptors for each category and level. Weir's (1990) rubric only allowed the raters to assign scores that ranged from 1 to 3 and the descriptors for each category included vague wording, such as "some" or "frequent." The inter-rater reliability coefficients for each of the categories in the analytic rubric were not very high, with one exception. The coefficient for spelling was higher than .700 for both exams. The coefficients ranged from .400 to .595 for content, organization, and cohesion, and from .120 to .352 for punctuation, grammar, and vocabulary. The coefficient for grammar was particularly low (.012). The low rater reliability coefficients may be a result of the few hours that the raters and I spent training and norming with the rubric, as well as the lack of calibration before each scoring session. RM had a considerably higher intra-rater coefficient for the TW exam (.862 compared to .352) and RK's coefficients showed the same trend (.552 for the PBTW and .642 for the TW exam). The raters' lack of familiarity and experience rating process-based exams may explain the lower reliability coefficient for the PBTW exam. Another reason for the lower reliability coefficient could be related to the test takers' inability to integrate sources, which could have been distracting for the raters. Although the rubric did not include a category for source integration, it seemed as though the raters attended to the test takers' lack of ability to integrate sources nonetheless. Finally, the scores that the participants received in the TW and PBTW exams correlated only moderately, suggesting that the exams measure different constructs. The PBTW exam measures, in addition to writing, reading, listening, note-taking skills, as well as the ability to integrate sources, all of which are important academic writing skills that students need to succeed in their classes.

6.1 Pedagogical implications

The main pedagogical implication of this study is that the process-based exam, when compared to impromptu timed-writing exams, might be a better fit for ESL academic writing programs that teach students planning strategies, synthesis, critical thinking, and source integration because this exam supports and validates these skills, unlike impromptu timed-writing exams. It is extremely important to consider content validity when designing a test (Carr, 2011; McNamara & Roever, 2006; Weigle, 2002). A test, whether its purpose is placement or achievement, should measure the curriculum upon which it is based (Carr, 2011). If the purpose of the writing exam is to check student achievement, and the goal of the course is to prepare students for regular university classes, then there is no doubt that the process-based writing exam has more content validity because impromptu timed-writing exams do not allow learners to use skills such as planning, synthesis, and source integration, all of which are skills that writing experts, many ESL textbooks, and research seem to support as crucial skills for university students.

Two other findings of this study that bear important pedagogical implications are the participants' and raters' preferences for process-based writing exams. The participants seemed to value the process through which the PBTW exam allows them to go. Many participants mentioned that they liked having planning time and learning about the topic through the source materials and these steps may have contributed to their higher scores in content. Test takers' opinions should be taken into consideration when designing tests because, as one of the main stakeholders, test takers may be more invested and motivated when they like the test that they are taking. Although the raters seemed to have difficulties scoring the PBTW exams and although

the rater reliability was lower for these exams, the raters also agreed that process-based writing exam have more content validity for ESL academic writing classes. Test designers should, however, ensure that they select rubrics with a wide range of possible scores and precise descriptors for each score level and category to obtain high rater reliability. In addition, test designers should apply rigorous rater training and norming, as well as sample essays for calibration.

Finally, one important aspect of classroom assessment is the issue of washback. If impromptu timed-writing exams continue to be administered in academic ESL writing classes, teachers will continue to teach students strategies to take such exams and spend class time training students for timed writing instead of focusing on preparing students to succeed in regular academic classes. Carr (2011) wrote that planning "tests that seem likely to cause positive washback is important because teachers will wind up teaching to the test" (p. 55). If process-based exams are implemented, teachers will focus more on teaching writing as a process and teaching students planning strategies, synthesis and source integration, among other skills that are part of the academic writing construct discussed above.

6.2 Limitations

The present study has limitations, including the population of the study and the sample size. As with any other study, the population in this study is not exactly the same as the population in all other ESL academic writing classes in the United States. The great majority of the participants were either from Brazil or China. Other ESL programs may have a different student population whose performance or perceptions of the TW and PBTW exams may not have been the same. Chinese students are used to preparing for and taking impromptu timed-writing

exams because of the university entrance exam. The Brazilian participants, on the other hand, are quite used to doing integrated reading and writing tasks, as they spend three years of high school preparing for the reading-to-write task that they take in the university entrance exams. This in turn could have explained the majority of the Brazilian students' preference for the PBTW exam. Although there were no significant differences between how the Brazilians performed in the two exams, there seemed to be a trend that indicated higher scores for the PBTW exam. The opposite was the case for the Chinese participants. There seemed to be a trend of higher scores for the TW exam, but no significant differences were found. The participants' number of years of formal instruction in English varied considerably, from seven months to seven years. In addition, their length of residence in the United States also varied greatly, from four months to three years. Although the participants took a placement exam and were placed in academic writing classes, their proficiency levels may have varied markedly because of the difference in the number of years of formal instruction and length of residence. Unfortunately I was unable to obtain any other measure of the participants' English proficiency to ensure that they were at a similar proficiency level. Even if I had asked the participants for their TOEFL scores, some participants had taken the exam months prior to data collection and their scores would not have been an accurate proficiency measure.

Another limitation of this study was the sample size. Although the present study had more participants than some of the studies reported in the literature review, such as Ellis and Yuan (2004) and David (under review), eighty-one participants is still not enough to make broad generalizations. Some aspects of the TW and PBTW exams were difficult to control and could have affected the results. During one of the semi-structured interviews, I learned that some of the participants who were taking listening and speaking classes had been discussing the issue of gun

control in their classes. They most likely had much more knowledge of the topic than the students who were not taking the listening and speaking course, and as discussed above, topic familiarity can have an effect on students' performance.

Larson-Hall (2009) and other statisticians suggested using the Bonferroni adjustment when performing multiple t tests and that is what I did with my data. However, some statisticians argue that using the Bonferroni adjustment might be too conservative, and instead suggest other types of corrections (Herrington, 2002). If that is indeed the case, then some of the results may have been different, which is another limitation of this study. O'Keefe (2003), for example, warned that the Bonferroni adjustment reduces statistical power dramatically and that researchers do not apply such alpha level corrections very consistently. He concluded that alpha level corrections should not be considered when researchers perform multiple *t* tests on the same set of data. However, he did suggest the use of alpha level adjustments if the results of the study are used to make important decisions. He explained that if that is the case, researchers might decide to use alpha level adjustments to decrease the probability of the results occurring by chance. If decisions have to be made about implementing a process-based writing exam in an ESL academic writing course, perhaps a lower alpha level might be justifiable.

The rubric was a limitation of this study, a fact which can be evidenced by the raters' comments about the rubric and the low intra-rater reliability coefficients. A rubric with more range and clearer descriptors would have been more appropriate for this study. There may not have been enough norming sessions with the raters, another possible explanation for the low intra-rater reliability. The raters would have also benefitted from reading some essays for the purpose of calibration before each scoring session. Moreover, it was impossible to ensure that the raters did not know which test type (TW or PBTW exam) they were scoring because the test

takers used ideas from the videos and article and cited them, as they were instructed to do. Knowing which test type they were scoring could have affected how the raters scored each exam. It was clear, for example, that the raters were attending to source integration even though the rubric did not include any category for how the test takers used sources. One example of this was RM's multiple comments about the test takers' lack of ability to integrate sources and the comment that she made about how distracting it was when they could not integrate sources well. Perhaps RM was compensating for the fact that the rubric did not include source integration by punishing the test takers' poor source integration skills when assigning scores for other categories, such as content or organization. If that was indeed the case, then some test takers might have received higher scores for content or organization had they been able to integrate the source materials successfully.

Finally, it is a common practice to use a third rater when the scores that two raters assign are more than 1 or 2 points apart. However, instead of using a third rater, I decided to average the scores because only 4% of the scores assigned for each category differed by more than 2 points. Had I asked a third rater to score the essays that differed by more than 2 points the results might have been different.

6.3 Future research

Integrated tasks and process-based exams have many promising research areas because there is still much researchers do not know about them. One area of research that deserves much attention is the comparison between lower and higher proficiency learners' performance on process-based exams. Many researchers claim that learners have to be more advanced in order to manage the complexities of such tasks (e.g., Cumming, 2013; Gebril & Plakans, 2013; Johnson

et al., 2012). However, to my knowledge, no one to date has investigated this issue with process-based exams. One study that investigated how learners of different levels perform in independent versus integrated tasks was Cumming et al.'s (2005), but they mainly investigated how learners of different proficiency levels deal with source materials. Gebril and Plakans (2013), for example, investigated how students of different levels differed in terms of fluency, syntactic and lexical complexity, grammatical accuracy, and source use when doing integrated tasks. In a study using TOEFL's integrated writing tasks (www.toefl.org), Sawaki et al. (2013) teased apart three skills that differentiate learners below and above the level of proficiency required for university admission. However, these studies did not compare how lower level students and higher level students perform in impromptu TW exams compared to how they perform in integrated tasks. In addition, the tasks used in these studies did not include group discussions or planning time.

More research is also needed to gather more information about test takers' opinions of the exams that they are required to take, as they are important stakeholders who often spend months preparing for a test and whose futures depend on how well they perform in them. Indeed, the ILTA Code of Ethics (2000) mandated such reflection and research because testers have a responsibility to understand the consequences of their tests on all stakeholders (see pages 6 and 7). Test developers may need to understand whether integrated tasks and process-based exams differentially affect students from different linguistic and cultural backgrounds. Do students from different cultures and native languages perform differently when they take integrated tests or PBTW exams? Brazilian students, for example, are most likely used to reading-to-write tasks because of the university entrance exams. The reading-to-write task in the Brazilian university entrance exams carries much weight and students spend three years of high school preparing for it. It would be interesting to investigate the processes that L2 writers from different backgrounds

go through when they take process-based writing exams and how they view and tackle each task involved in the exam.

Teachers are also important stakeholders for some tests, especially classroom-based assessments. They are often taken for granted and not included in the test design process. Researchers should, therefore, also investigate and consider teachers' perceptions of exams that oftentimes affect the way they teach and what they teach in their classrooms.

6.4 Summary

Process-based writing exams may be a better way to evaluate academic writing when compared to impromptu TW exams because processed-based writing may better match the construct of academic writing than TW exams do. But implementing process-based writing exams might not be easy. They take more time to administer. On the other hand, process-based writing exams encourage students to view writing as a process, not a product, because the exams provide learners with the opportunity to discuss the topic and learn about the topic, as well as additional time to plan their writing. Process-based exams may also help students write essays with higher quality content and more sophisticated vocabulary because the test takers can use ideas from the readings, videos, and discussions. And they can use vocabulary from the source materials. In addition, process-based writing exams may have more content validity for academic writing classes in which source integration plays a big role. Impromptu TW exams may not allow students to demonstrate the skills acquired through academic writing classes. In such classes they learn planning, synthesis, source integration, and so on, all of which are skills that should be part of the construct of academic writing. Another reason why process-based exams should be used in the place of impromptu TW exams when students' achievement is being

assessed is the fact that students enjoy the process through which they go when they take such exams. They value the discussion, source materials, and planning time, and they seem to believe that these contribute to their success in the writing task. When implementing process-based exams, however, test designers must give careful consideration to the rubric which will be used to evaluate learners' writing. Test designers must use rubrics that have a wide range of possible scores and clear descriptors for each score band and category. Raters must be highly trained, and a robust process of double rating must be implemented (McNamara & Roever, 2006, p. 27). Such measures will help control the effects that the complex, process-based task components or the rater may have on scores.

APPENDICES

APPENDIX A: Videos

Videos from obesity prompt:

http://abcnews.go.com/Health/video/large-sugary-drink-ban-passes-new-york-city-17227911
http://abcnews.go.com/Nightline/video/mcdonalds-calorie-counts-nyc-big-soda-ban-17232674

Videos for gun control prompt:

http://www.youtube.com/watch?v=vtAAI4xnmzE
http://abcnews.go.com/WNT/video/aurora-colorado-shooting-gun-control-laws-16829309

APPENDIX B: Reading passages

Article for obesity prompt:

http://www.nytimes.com/2011/07/24/opinion/sunday/24bittman.html
http://www.nytimes.com/2013/12/15/opinion/sunday/kristof-the-killer-who-supports-gun-control.html

APPENDIX C: Rubric

Rubric (Weir, 1990)

A. *Relevance and adequacy of content*
   0. The answer bears almost no relation to the task set. Totally inadequate answer.
   1. Answer of limited relevance to the task set. Possibly major gaps in treatment of topic and/or pointless repetition.
   2. For the most part answers the tasks set, though there may be some gaps or redundant information.
   3. Relevant and adequate answer to the task set.

B. *Compositional organization*
   0. No apparent organization of content.
   1. Very little organization of content. Underlying structure not sufficiently controlled.
   2. Some organizational skills in evidence, but not adequately controlled.
   3. Overall shape and internal pattern clear. Organizational skills adequately controlled.

C. *Cohesion*
   0. Cohesion almost totally absent. Writing so fragmentary that comprehension of the intended communication is virtually impossible.
   1. Unsatisfactory cohesion may cause difficulty in comprehension of most of the intended communication.
   2. For the most part satisfactory cohesion although occasional deficiencies may mean that certain parts of the communication are not always effective.
   3. Satisfactory use of cohesion resulting in effective communication.

D. *Adequacy of vocabulary for purpose*
   0. Vocabulary inadequate even for the most basic parts of the intended communication.
   1. Frequent inadequacies in vocabulary for the task. Perhaps frequent lexical inappropriacies and/or repetition.
   2. Some inadequacies in vocabulary for the task. Perhaps some lexical inappropriacies and/or circumlocution.
   3. Almost no inadequacies in vocabulary for the task. Only rare inappropriacies and/or circumlocution.

E. *Grammar*
   0. Almost no grammatical patterns inaccurate.
   1. Frequent grammatical inaccuracies.
   2. Some grammatical inaccuracies.
   3. Almost no grammatical inaccuracies.

F. *Mechanical accuracy I (punctuation)*
   0. Ignorance of conventions of punctuation.
   1. Low standard of accuracy in punctuation.

2. Some inaccuracies in punctuation.
3. Almost no inaccuracies in punctuation.

G. *Mechanical accuracy II (spelling)*
0. Almost all spelling inaccurate.
1. Low standard of accuracy in spelling.
2. Some inaccuracies in spelling.
3. Almost no inaccuracies in spelling.

APPENDIX D: Post-writing questionnaire

Post-writing Questionnaire Participant ID _____

For this research project, you took two different types of timed writing exams. One was shorter and required you to write an essay in 45 minutes. The other was longer and required you to watch two short videos, read one article, discuss the topic with your classmates, plan your essay, and then write it in 45 minutes. Please answer the questions about the two exams.

1. Which exam do you think was easier?

a) The shorter timed writing exam
b) The longer timed writing exam with the videos, lectures and discussion
c) They were equally easy/difficult (circle one)

2. Why did you think one exam was easier than the other? If you think they were equally easy/difficult, skip this question.

3. What did you think about the videos that you watched? You can choose more than one answer for this question.

a) They were easy to understand
b) They were difficult to understand
c) They helped me think of ideas for the essay
d) They did not help me think of ideas for the essay

4. What did you think about the article that you read? You can choose more than one answer for this question.

a) It was easy to read
b) It was difficult to read
c) It helped me think of ideas for the essay
d) It did not help me think of ideas for the essay

5. What did you think about the group discussion?

a) It helped me think of ideas for the essay
b) It did not help me think of ideas for the essay
c) It was not related to what I wrote in my essay

6. Did you use the videos to help you support your ideas in the essay?

a) Yes

b) No

7. Did you use the article to help you support your ideas in the essay?

a) Yes
b) No

8. Did you use the ideas that you discussed in your group in the essay?

a) Yes
b) No

9. If you did not use the videos, article or ideas in the group discussion, why did you choose not to do so?

10. What was difficult and/or easy about doing the shorter timed writing exam?

11. What was difficult and/or easy about doing the longer timed writing exam with the videos, article and discussion?

12. Which exam did you prefer taking?

a) The shorter timed writing exam with two topics
b) The longer timed writing exam with the videos, lectures and discussion

13. Why did you prefer taking that exam?

APPENDIX E: Semi-structured interview questions

Questions for the students:

1. What did you think of the two types of writing exams that you had to take? Let's begin with the shorter exam. What about the longer exam?

- Prompts for this question include:

What did you think of the time limit?
What did you think of the topics?
What did you think of the videos and article?
What did you think of the group discussion?
What did you think of the time for planning?

2. What did you like about the two exams?
3. What did you dislike about the two exams?
4. Did you think you did better at one exam than the other? Which exam? Why?
5. What are some of the problems that you face when taking a timed writing exam?
6. If you could choose a way to be evaluated for your writing skills, what evaluation method would you choose?
7. How can you prepare for taking a timed writing exam?
8. What can you learn from taking timed writing exams?

Questions for the raters:

1. What is your overall impression of the two exams?
2. What are the advantages and disadvantages of each exam?
3. What was it like to rate the essays for the two exams?
4. Were there any difficulties scoring the exams?
5. As both raters and ESL teachers, how representative of what you do in the classroom are the two exams

APPENDIX F: Guidelines for clauses

Guidelines for Clauses (Polio, 1997)

a. A clause equals an overt subject and a finite verb. The following are only one clause each:
   He left the house and drove away.
   He wanted John to leave the house.
b. Only an imperative does not require a subject to be considered a clause.
c. In a sentence that has a subject with only an auxiliary verb, do not count that subject and verb as a separate clause (e.g. John likes to ski and Mary does too; John likes to ski, doesn't he?; John is happy and Mary is too)

Error Guidelines

a. Do not count spelling errors (including word changes like "there/their").
b. Be conservative about counting comma errors; don't count missing commas between clauses or after prepositional phrases. Comma errors related to restrictive/non-restrictive relative clauses *should* be counted. Extraneous commas should also be considered errors.
c. Base tense/reference errors on preceding discourse; do not look at the sentence in isolation.
d. Don't count British usage as errors, (e.g. "in hospital," "at university," collective nouns as plural).
e. Be lenient about article errors from translations of proper nouns.
f. Don't count errors in capitalization.
g. Count errors that could be made by native speakers (e.g. between you and I).
h. Do not count register errors related to lexical choices (e.g. lots, kids).
i. Disregard an unfinished sentence at the end of the essay.

REFERENCES

REFERENCES

Abdel Latif, M. M. (2013). What do we mean by writing fluency and how can it be validly measured? *Applied Linguistics, 34*(1), 99-105.

Ai, H., & Lu, X. (2013). A corpus-based comparison of syntactic complexity in NNS and NS university students' writing. *Automatic Treatment and Analysis of Learner Corpus Data*, *59*.

Armstrong, K. M. (2010). Fluency, accuracy, and complexity in graded and ungraded writing. *Foreign Language Annals*, *43*(4), 690-702.

Baralt, M. (2012). Coding qualitative data. In A. Mackey & S. M. Gass (Eds.), Research *Methods in Second Language Acquisition: A Practical Guide* (pp. 95-116). Chichester, UK: John Wiley & Sons, Ltd.

Baralt, M., Gilabert, R., & Robinson, P. (2014). *Task sequencing and instructed second language learning*. London: Bloomsbury.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*(1), 54-74.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, N.J: L. Erlbaum Associates.

Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford: Oxford University Press.

Chaudron, C., & Parker, K. (1990). Discourse markedness and structural markedness: The acquisition of English noun phrases. *Studies in Second Language Acquisition, 12,* 43–64.

Cho, Y. (2001). Examining a process-oriented writing assessment in a large-scale ESL testing context. (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.

Chung, T. (2003) A corpus-comparison approach for term extraction. *Terminology 9,* 2: 221-246

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159.

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. Journal of Educational Measurement, 37(2), 163-178.

Connor-Linton, J. & Polio, C. (2014). Comparing perspectives on L2 writing: Multiple analyses of a common corpus: Introduction. *Journal of Second Language Writing, 23*, 1-9.

Cooper, A., & Bikowski, D. (2007). Writing at the graduate level: what tasks do professors actually require?. *Journal of English for Academic Purposes*, *6*(3), 206-221.

Creswell, J. W. & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research.* Thousand Oaks, CA: Sage.

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly, 10*(1), 1-8.

Cumming, A., Grant, L., Mulcahy-Ernt, P., & Powers, D. E. (2004). A teacher-verification study of speaking and writing prototype tasks for a new TOEFL. *Language Testing, 2,* 107-145.

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing, 10*(1), 5-43.

David, V. (under review). *A comparison of two methods of assessing L2 writing: Process-based and impromptu timed writing exams*. Manuscript submitted for publication.

Deane, P., Odendahl, N., Quinlan, T., Fowles, M., Welsh, C., & Bivens-Tatum, J. (2008). Cognitive models of writing: Writing proficiency as a complex integrated skill. *ETS Research Report Series, 2008*(2), i-36.

Delaney, Y. A. (2008). Investigating the reading-to-write construct. *Journal of English for academic purposes, 7,* 140-150.

Dollahite, N. E., & Haun, J. (2007). *Sourcework: Academic writing from sources*. Thomson/Heinle.

Duff, P. (2012). How to carry out case study research. In A. Mackey & S. M. Gass (Eds.), *Research Methods in Second Language Acquisition: A Practical Guide* (pp. 95-116). Chichester, UK: John Wiley & Sons, Ltd.

Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition, 26*(01), 59-84.

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

Esmaeili, H. (2002). Integrated reading and writing tasks and students' reading and writing performance in an English language test. *The Canadian Modern Language Review, 58,* 599-622.

Ferris, D. (2009). *Teaching college writing to diverse student populations*. Ann Arbor: University of Michigan Press.

Field, A. (2009). *Discovering statistics using SPSS*. Thousand Oaks, CA: Sage publications.

Flower, L. S. and J. R. Hayes. 1981. 'The pregnant pause: an inquiry into the nature of planning,' Research in the Teaching of English 15/3: 229–43

Fritz, E., & Ruegg, R. (2013). Rater sensitivity to lexical accuracy, sophistication and range when assessing writing. *Assessing Writing, 18*(2), 173-181.

Gass, S. M., & Selinker, L. (2008). Second language acquisition: An introductory course (3rd ed.). New York: Routledge/Taylor and Francis Group.

Gebril, A. (2010). Bringing reading-to-write and writing-only assessment tasks together: A generalizability analysis. *Assessing Writing, 15*(2), 100-117.

Gebril, A., & Plakans, L. (2013). Toward a transparent construct of reading-to-write tasks: The interface between discourse features and proficiency. *Language Assessment Quarterly, 10*(1), 9-27.

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

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgements of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing, 18,* 218-238.

Hale, G., Taylor, C., Bridgeman, B., Carson, J., Kroll, B., & Kantor, R. (1996). *A study of writing tasks assigned in academic degree programs*. Princeton, NJ: Educational Testing Service.

Harley, B., & King, M. L. (1989). Verb lexis in the written compositions of young L2 learners. *Studies in Second Language Acquisition, 11*, 415–440.

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

Hayes, J. R., & Flower, L. (1980). Identifying the organization of writing processes. In L. W. Gregg and E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 31-50). Hillsdale, NJ: Lawrence Erlbaum Associates.

He, L., & Shi, L. (2008). ESL students' perceptions and experiences of standardized English writing tests. *Assessing Writing, 13*(2), 130-149.

He, L., & Shi, L. (2012). Topical knowledge and ESL writing. *Language Testing, 29*(3), 443-464.

Heatley, A., Nation, I.S.P. and Coxhead, A. (2002). RANGE and FREQUENCY programs. http://www.vuw.ac.nz/lals/staff/Paul_Nation

Herrington, R. (2002). Controlling the false discovery rate in multiple hypothesis testing. On www.unt.edu/benchmarks/archives/2002/april02.rss.htm. Research and Statistical Support web site. University of North Texas, Denton

Horowitz, D. M. (1986). What professors actually require: Academic tasks for the ESL classroom. *TESOL Quarterly*, *20*(3), 445-462.

Howell, D. C. (2002). *Statistical methods for psychology*. Pacific Grove, CA: Duxbury/Thomson Learning.

Hughes, A. (2003). *Testing for language teachers*. Cambridge; New York: Cambridge University Press.

Hunt, K. W. (1965). Grammatical Structures Written at Three Grade Levels. NCTE Research Report No. 3.

Hyltenstam, K. (1988). Lexical characteristics of near-native second-language learners of Swedish. *Journal of Multilingual and Multicultural Development, 9*, 67–84.

ILTA. (2000). *ILTA code of ethics*. Available at http://www.iltaonline.com/images/pdfs/ilta_code.pdf

Jackson, D. O., & Suethanapornkul, S. (2013). The cognition hypothesis: A synthesis and meta-analysis of research on second language task complexity. *Language Learning, 63*(2), 330-367.

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

Johnson, M. D., Mercado, L., & Acevedo, A. (2012). The effect of planning sub-processes on L2 writing fluency, grammatical complexity, and lexical complexity. *Journal of Second Language Writing, 21*(3), 264-282.

Kellogg, R. T. (1988). Attentional overload and writing performance: Effects of rough draft and outline strategies. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(2), 355-365.

Kormos, J. (2011). Task complexity and linguistic and discourse features of narrative writing performance. *Journal of Second Language Writing*, *20*(2), 148-161.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*(2), 275-304.

Kuiken, F., Mos, M., & Vedder, I. (2005). Cognitive task complexity and second language writing performance. *Eurosla yearbook, 5*(1), 195-222.

Kuiken, F., & Vedder, I. (2008). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing*, *17*(1), 48-60.

Larson-Hall, J. (2009). *A guide to doing statistics in second language research using SPSS*. Routledge.

Laufer, B. (1994).The lexical profile of second language writing: Does it change over time? *RELC Journal*, *25*, 21–33.

Lee, Y. J. (2006). The process-oriented ESL writing assessment: Promises and challenges. *Journal of Second Language Writing*, *15*(4), 307-330.

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*(1), 34-50.

Leki, I. (1991). A new approach to advanced ESL placement testing. Writing Program Administration, 14 (3), 53–68.

Linnarud, M. (1986). *Lexis in composition: A performance analysis of Swedish learners' written English*. Lund, Sweden: CWK Gleerup.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474-496.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, *45*(1), 36-62.

Lu, X. (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *Modern Language Journal*, *96*(2), 190-208.

Lumley, T., & McNamara, T. F. (1993). Rater characteristics and rater bias: Implications for training.

Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. Evaluation & the Health Professions, 13(4), 425-444.

Malvern, D. D., & Richards, B. J. (1997). A new measure of lexical diversity. In A. Ryan & A. Wray (Eds.), *Evolving models of language* (pp. 58–71). Clevedon: Multilingual Matters.

McCarthy, P.M. (2005). *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity* (MTLD) (Unpublished doctoral dissertation). University of Memphis, Memphis, TN.

McNamara, D. S., Louwerse, M. M., Cai, Z., & Graesser, A. (2005, January 1). Coh-Metrix

version 1.4.

McNamara, T., & Roever, C. (2006) *Language testing: The social dimension.* Malden, MA: Blackwell Publishing.

Morse, J. M. (1991). Approaches to qualitative-quantitative methodological triangulation. *Nursing Research, 40,* 120-123.

Nation, I. S. P. (2005). Range and frequency: Programs for Windows based PCs [Computer software and manual]. Retrieved from http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards*. Washington, DC: Authors.

O'keefe, D. J. (2003). Colloquy: Should familywise alpha be adjusted? Against familywise alpha adjustment. *Human Communication Research, 29*(3), 431-447.

Ong, J., & Zhang, L. J. (2013). Effects of task complexity on the fluency and lexical complexity in EFL students' argumentative writing. *Journal of Second Language Writing, 19*(4), 218-233.

Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing, 13*(2), 111-129.

Polio, C. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning, 47,* 101-143.

Polio, C. & Shea, M. (2014). Another look at accuracy in second language writing development. *Journal of Second Language Writing, 23*, 10-27.

Polio, C. & Yoon, H.J. (2014). A longitudinal study of written language development in two genres. Second Language Writing Symposium, Arizona State University, Tempe, AZ. November 2014.

Prior, P. (1998). Writing/disciplinarity: A sociohistoric account of literate activity in the academy. Mahwah, NJ: Lawrence Erlbaum.

Powers, D. E., & Fowles, M. E. (1999). Test-takers' judgments of essay prompts: Perceptions and performance. *Educational Assessment, 6*(1), 3-22.

Rea-Dickins, P. (1997). So, why do we need relationships with stakeholders in language testing? A view from the UK. *Language Testing, 14*(3), 304-314.

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 15*(1), 18-39.

Richards, K. (2003). *Qualitative inquiry in TESOL*. Basingstoke: Palgrave Macmillan.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, *22*(1), 27-57.

Sawaki, Y., Quinlan, T., & Lee, Y. W. (2013). Understanding learner strengths and weaknesses: assessing performance on an integrated writing task. *Language Assessment Quarterly*, *10*(1), 73-95.

Shi, L. (1998). Effects of prewriting discussions on adult ESL students' compositions. *Journal of Second Language Writing*, *7*(3), 319-345.

Tedick, D. J. (1990). ESL writing assessment: Subject-matter knowledge and its impact on performance. *English for Specific Purposes*, *9*(2), 123-143.

Way, P., Joiner, E. G., & Seaman, M. (2000). Writing in the secondary foreign language classroom: The effects of prompts and tasks on novice learners of French. *Modern Language Journal, 84*(2), 171-184.

Weigle, S. C. (1998). Using FACETS to model rater training effects. Language Testing, 15(2), 263-287.

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

Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing, 9,* 27-55.

Weir, C. J. (1990). *Communicative language testing.* NJ: Prentice Hall Regents.

Wigglesworth, G., & Storch, N. (2009). Pair versus individual writing: Effects on fluency, complexity and accuracy. *Language Testing*, *26*(3), 445-466.

Winfield, F. E., & Barnes-Felfeli, P. (1982). The effects of familiar and unfamiliar cultural context on foreign language composition. *The Modern Language Journal*, *66*(4), 373-378.

Winke, P. (2013). The effectiveness of interactive group orals for placement testing. In K. McDonough & A. Mackey (Eds.), *Second language interaction in diverse educational contexts* (pp. 247-268). John Benjamins Publishing Company.

Winke, P., & Lim, H. (2015). ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study. *Assessing Writing, 25*(2), 37-53.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). Second language development in writing: Measures of fluency, accuracy, and complexity (Report No. 17). Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center.

Worden, D. L. (2009). Finding process in product: Prewriting and revision in timed essay responses. *Assessing Writing*, *14*(3), 157-177.

Yigitoglu, N. (2008). *A pathway between academic and ESL classes: Academic tasks and their potential impact on teaching and testing writing. (*Unpublished Master's Thesis). Michigan State University.