The Effect of the Prompt on Writing Product and Process: A Mixed Methods Approach

Mark Derek Chapman

THE EFFECT OF THE PROMPT ON WRITING PRODUCT AND PROCESS: A MIXED METHODS APPROACH

MARK DEREK CHAPMAN

Ph.D

2016

UNIVERSITY OF BEDFORDSHIRE

THE EFFECT OF THE PROMPT ON WRITING PRODUCT AND PROCESS: A MIXED METHODS APPROACH

by

MARK DEREK CHAPMAN

A thesis submitted to the University of Bedfordshire in partial fulfilment of the requirements for the degree of Doctor of Philosophy

2016

THE EFFECT OF THE PROMPT ON WRITING PRODUCT AND PROCESS: A MIXED METHODS APPROACH

MARK DEREK CHAPMAN

ABSTRACT

The aim of this thesis is to investigate the effect of the writing prompt on test takers in terms of their test taking processes and the final written product in a second language writing assessment context. The study employs a mixed methods approach, with a quantitative and a qualitative strand. The quantitative study focuses on an analysis of the responses to six different writing prompts, with the responses being analyzed for significant differences in a range of key textual features, such as syntactic complexity, lexical sophistication, fluency and cohesion. The qualitative study incorporates stimulated recall interviews with test takers to learn about the aspects of the writing prompt that can have an effect on test taking processes, such as selecting a prompt, planning a response, and composing a response.

The results of the quantitative study indicate that characteristics of the writing prompt (domain, response mode, focus, number of rhetorical cues) have an effect on numerous textual features of the response; for example, fluency, syntactic complexity, lexical sophistication, and cohesion. The qualitative results indicate that similar characteristics of the writing prompt can have an effect on how test takers select a prompt, and that the test time constraint interacts with the prompt characteristics to affect how test takers plan and compose their responses. The topic and the number of rhetorical cues are the prompt characteristics that have the greatest effect on test taking processes.

The main conclusion drawn from the study findings are that several prompt characteristics should be controlled if prompts are to be considered equivalent. Without controlling certain prompt characteristics, both test taking processes and the written product will vary as a result of the prompt. The findings raise some serious questions regarding the inferences that may legitimately be drawn from writing scores. The findings provide clear guidance on prompt characteristics that should be controlled to help ensure that prompts present an equivalent challenge and opportunity to test takers to demonstrate their writing proficiency. This thesis makes an original contribution to the second language writing assessment literature in the detailed understanding of the relationships between specific prompt characteristics and textual features of the response.

ACKNOWLEDGMENTS

This thesis is the result of input and advice from my supervisors, colleagues, and friends. I owe particular gratitude to my PhD supervisors, Professor Liz Hamp-Lyons and Professor Tony Green. I know I have been very fortunate to have access to Liz's vast knowledge of the second language writing assessment field. I could not have hoped for a better supervisor for the work I have undertaken for my PhD. Tony's guidance with the statistical handling of the data in my thesis was invaluable and his views on my writing and ideas were always gratefully received. In addition, I would like to thank Professor Roger Murphy and Dr. Julio Gimenez for their feedback on earlier iterations of my research.

The motivation for this thesis came from my time at The English Language Institute (ELI) at The University of Michigan. I am grateful to Dr. Gary Buck who brought me to the ELI and to all my former colleagues there for the incredibly creative and stimulating environment they forged. The test development team at the ELI of Robin, Stephen, Fabiana, and later Rachele provided the genesis for my research and my career in language testing. Dr. Jayanti Banerjee was extremely influential as I developed and narrowed the focus of my work. She generously gave her time to help bring structure to my thoughts and I greatly appreciate her insightful input and insistence that I pursue original ideas. I also owe a large thankyou to Dr. Spiros Papageorgiou for his interest in my research and willingness to provide advice on the research methods. The staff at the Center for Statistical Consultation and Research at The University of Michigan were very important in guiding my understanding of how to analyze the data I collected.

My colleagues at Michigan who worked with me on the MELAB Writing program provided me with great support as I formulated hypotheses about writing prompts and how they influenced the responses we spent many hours reading and scoring. In particular, I would like to thank Sarah Goodwin, Caitlin Gdowski, Aaron Ohlrogge, Gad Lim, Ummehaany Jameel, Heather Elliott, and Crystal Collins. This thesis would never have come about without the help of Becky Orta who was patience personified as she helped me identify hundreds of MELAB essays. Also, I owe much gratitude to Dr. Georgia Wilder of The University of Toronto who allowed me regular access to MELAB test administrations and test takers at the Toronto MELAB test center. Thanks also to the many test takers who agreed to talk with me about their experiences of taking the MELAB Writing test. Hearing the opinions of the individuals who take high-stakes tests is an experience no language testing researcher should go without.

Of course, I must also thank my family for all the times I was too busy working on this thesis to be with them. I am sure there were more times than they care to remember. Thanks to my wife and daughter for their patience and understanding and to my parents for their constant support and encouragement.

Table of Contents

List of Tables

List of Figures

Chapter 1 – Introduction

Anyone who has ever taken an important examination understands the need that it be fair and for the scores awarded to be reliable. Some examinations can have serious consequences for the candidates' lives. Examples of high-stakes decisions that can involve the requirement to pass a test or obtain a particular score are admission to higher education, professional certification or licensure, and even migration to another country. When test scores can have such serious consequences, it is essential that these tests are well designed and reliably scored.

However, the process of designing a test that consistently produces reliable scores is not simple. For the test to be secure, the questions and tasks that appear on the test must be unknown to the test takers. If test takers know the specific test content in advance, they can prepare for it and the resulting score will no longer be a realistic interpretation of their ability level. Hence, for a test to be secure and the scores to be meaningful, it is necessary for each new test form to contain different tasks and questions from those of previous forms. The issue that then arises is how to guarantee that scores awarded to different test forms are consistent over time. How can we be sure that a B+ or a particular numeric score on this year's test means the same as on last year's? This issue of consistent score interpretation is at the heart of test reliability.

A challenging problem for test designers is ensuring that constructed response tasks (ones that require test takers to produce a written or spoken response) are of equal difficulty and elicit responses that can be consistently awarded reliable scores. It is particularly important to ensure that test takers are not advantaged or disadvantaged, in terms of the score they are awarded, by the tasks they select to respond to. The issue of the reliable measurement of constructed response tasks has troubled those involved with testing issues, for over a hundred years (Edgeworth, 1890; Ruch, 1929; Hartog & Rhodes, 1936; Stalnaker, 1937; Diederich, 1946; Wiseman, 1949; Kincaid, 1953). Much of this historical work addressed concerns over the reliability of the direct assessment of writing proficiency. Particularly in the United States, human ratings of written language proficiency were seen as subjective and unreliable, with preference being expressed for the indirect assessment of writing proficiency via multiple choice items that evaluated skills thought to underlie the construct of writing proficiency, such as spelling, grammatical accuracy, and vocabulary choice. These concerns with rater reliability were typified in the work of Diederich, French, & Carlton (1961) that questioned the capacity of human raters to reliably evaluate written language in the academic context and led to the virtual disappearance of the direct assessment of writing for many years. In the UK, the concerns over the reliable rating of written language were less prevalent than in the US (Hamp-Lyons, 2002) with the validity and washback of the test being viewed as more important than the reliability of the human awarded scores.

The validity of writing scores and the inferences about writing proficiency that may be drawn from these scores have been questioned in the second language writing assessment literature (Chapelle, 1999; Lumley, 2005; Shaw & Weir, 2007; Weir, 2005). The prompts used to elicit written responses are one potential source of measurement error in second language writing assessment, as Chapter 2 shows. Prompts may introduce measurement error into the assessment process by being biased (systematically disadvantaging at least one subset of the test population) (Ruth & Murphy, 1984; Hamp-Lyons, 1988;

Johns, 1991; Kroll & Reid, 1994; He & Shi, 2008), by varying significantly in difficulty (Breland, Lee, Najarian, & Muraki, 2004; Petersen, 2009; Polio & Glew, 1996; Tavakoli, 2009), and by exhibiting variability in the linguistic features of the responses elicited (Cumming, Kantor, Baba, Erdosy, Keanre, & James, 2005; Greenberg, 1981; O'Loughlin & Wigglesworth, 2007; Ong & Zhang, 2010). These issues can cause problems with score reliability and with score interpretation.

This study focuses on the effects of the writing prompts presented to test takers on one particular high-stakes assessments of writing proficiency in order to illuminate the issues of how the prompt may have an effect on both test taking processes and the written products elicited. The goal of test developers is that different writing prompts that appear on multiple test forms will not have any effect on test takers and test performance, so all prompts may be considered equivalent. However, establishing the equivalence of writing prompts and, indeed what equivalence means is not a simple task. If a writing assessment is to be fair for each test taker, and test takers are to be provided with prompts that provide equal opportunities to all test takers, the prompts administered within a single test form and across multiple test forms must be equivalent in:

1. The opportunities they present to test takers to demonstrate their true level of writing proficiency, and;
2. The scores that are awarded to responses to the prompts.

If writing prompts are to be considered equivalent they should elicit samples of written language that are easily comparable: That is, the responses to different prompts should not be identical but they should not differ significantly in length or in the complexity, sophistication, and accuracy of the written language produced, for test takers of similar levels of writing proficiency. There is also the possibility that the scores awarded to responses that differ significantly by certain textual features may also show significant variation. If scores differ significantly by the prompt that is responded to, questions must be raised about the fairness and validity of the test because the score a test taker receives may depend, to some extent on the prompt that was selected and responded to.

The broad questions to be addressed in this study are:

> Do the writing prompts presented on one high-stakes writing test offer all individuals who are required to take the test an equal opportunity to demonstrate their true level of writing proficiency?
>
> If writing prompts are not equivalent, what are the characteristics of writing prompts that contribute to the lack of equivalence?
>
> What is the effect of writing prompts that are not equivalent on the responses written to such prompts and on the test takers' interactions with the writing prompts?

This research project is situated within the assessment context of a specific examination, the *Michigan English Language Assessment Battery*, or MELAB. The MELAB is a high-intermediate to advanced level English language proficiency test that is used for admissions and professional licensure purposes. The MELAB was originally developed by the English Language Institute of the University of Michigan and is

now maintained and administered by Cambridge Michigan Language Assessments. Independent reviews of the MELAB are available in the language testing literature (Chalhoub-Deville, 2003; Purpura, 2005; Weigle, 2000) and a Technical Review (CaMLA, 2014) is also available.

Chapters 2 and 3 together present a review of literature relevant to the aims of this project, the effect of the writing prompt on test taking processes and written product. Chapter 2 reviews the literature on prompt effect from two perspectives. First, the literature on the effect of writing prompts on the written product is reviewed. This incorporates the effect of the prompt on both the score awarded to the written product and the effect of the prompt on the textual features of the response. Second, the literature on the effect of writing prompts on the test taking process is reviewed. Chapter 3 presents a review of literature that focuses on the distinguishing characteristics of writing prompts. Previous research has reported on categorization frameworks for writing prompts in an attempt to describe the characteristics of prompts that may be used to distinguish between different types of prompts. The development of a prompt categorization framework was an important foundation for this project as it guided the identification of the prompts that would be analyzed. This work is reported in Chapter 4.

Following the review of literature, the research questions that are addressed in this study are presented in Chapter 4. Chapter 4 then describes the materials and methods employed for the quantitative side of the study. This includes an account of how the prompt categorization framework was developed, based on the literature review from Chapter 3 and also from a range of interviews with raters and test takers who provided guidance on which prompt characteristics best distinguish between different writing prompts. After this categorization framework has been described, Chapter 4 also details the six writing prompts that were identified as the ones that would be analyzed for this project. Finally, in Chapter 4 the process of identifying the textual features used to analyze the written products is described.

Modern computational linguistic tools are helpful for investigating these issues. In the past, the writing assessment field had relied on variance in scores awarded to written products to determine the effect of the prompt on the response (Hoetker, 1982; Hoetker & Brossell, 1989; Leu, Keech, Murphy, & Kinzer, 1982). More recently, however researchers have been able to take advantage of language corpora and computational linguistic tools such as Coh-metrix (Graesser, McNamara, Louwerse, & Cai, 2004) and the Corpus of Contemporary America English (Davies, 2008). These tools were used to analyze a range of textual features in responses to several different writing prompts. The resulting data were analyzed to determine whether there were significant differences in the textual features of responses to the different writing prompts.

In Chapter 5 the results of the quantitative study are laid out. The results provide evidence as to whether the writing prompts elicit written language that is stable and consistent across responses to different prompts. The statistical analyses applied reveal whether there are significant differences in the textual features of the responses to the different prompts. These results are then discussed in Chapter 6 in which, multiple examples of test taker writing are presented to exemplify the significant differences in textual features reported in Chapter 5.

This quantitative approach is complemented by a qualitative approach, described in Chapter 7 to further investigate the effect of writing prompts on test takers and test taking processes. In order to learn more about the effect of the writing prompt, test takers were interviewed, using a stimulated recall approach after they had completed a writing test. The purpose of these interviews is to have test takers recall how they interact with the writing prompt as they choose a prompt to respond to and then compose a response under test conditions. The qualitative methodology and the results are both presented in Chapter 7 as the qualitative study plays a supporting role in the thesis to the quantitative study described in the previous chapters.

Chapter 8 synthesizes the key findings from the quantitative and qualitative approaches. Finally, in Chapter 9 answers to the research questions are restated and implications from the findings are presented. Future research directions that arise from the findings are also proposed.

In summary, this project adopts a mixed methods approach to investigate how writing prompts may have an effect on the written product elicited by the prompt, in terms of the textual features of the response and score awarded, and on the process of producing that response as the test taker interacts with the prompt during a writing test.

Chapter 2 – Review of Literature (1)

In this chapter I review the literature on the effect of writing prompts on the written product elicited and on the test taking process.

I consider the effects of prompts on written *products* in terms of:

i) the scores awarded, and
ii) key textual features of the response.

In relation to effects of the prompt on the test taking *process*, I consider:

i) what influences the test takers' choice of prompt, and
ii) how test takers interact with the prompt during the test.

This chapter will also review the literature relating to a range of variables (test taker factors and examination context factors) that may interact with how the prompt influences performance on a writing test. It is important to understand these potentially confounding variables so their impact may be minimized in the research design. This chapter will also establish how the current work makes an original contribution to the literature on writing prompt effect.

2.1 Effect of prompt on test scores

Many early studies into the relationship between writing prompt characteristics and written product used holistic scores awarded to the response to determine the effect of different prompt characteristics. Typically, these studies found that there was no statistically significant relationship between changes in prompt characteristics and the score awarded to the response, (Woodworth & Keech, 1980; Greenberg, 1981; Leu, Keech, Murphy & Kinzer, 1982; Brossell & Ash, 1984; Hoetker & Brossell, 1989). This early work, focusing on how holistic scores varied, indicated that the prompt did not have any significant effect on the written product. A description of some of these studies, their populations, prompt characteristics investigated, and outcomes will be given below.

2.1.1 Effect of prompt characteristics on holistic scores

Brossell (1983) administered three writing prompts to 360 undergraduate education majors at Florida State University. Six essay topics were used and each topic was written in three levels of "information load," (p.166). The prompts varied in length and complexity, as shown in the example below.

Level 1*: Violence in the schools.*

Level 2*: According to recent reports in the news media, there has been a marked increase in incidents of violence in public schools. Why, in your view, does such violence occur?*

Level 3*: You are a member of a local school council made up of teachers and citizens. A recent increase in incidents of violence in the schools has gotten widespread coverage in the local news media. As a teacher, you are aware of the problem, though you have not been personally involved in an accident. At*

its next meeting, the council elects to take some action. It asks each member to draft a statement setting forth his or her views on why such violence occurs. The statements will be published in the local newspaper. Write that statement expressing your own personal views on the causes of violence in the schools.

Participants were given 45 minutes to complete one essay under test conditions and each response was holistically rated on a four-point scale. The results showed no statistically significant effects of either topic or level of rhetorical specification on the overall score awarded. The author reported that the degree of rhetorical specification had a greater effect on the score than did the topic, even though neither was statistically significant. Essays written in response to the full rhetorical specification prompt had the lowest mean scores and shortest mean length compared to the other two levels of prompts. The author concluded that "the effects of full rhetorical contexts, especially in a scenario form, can be counterproductive to writers by inducing them to repeat needlessly the information in a topic and thus waste time," (p.172).

Brossell & Ash (1984) examined two specific prompt characteristics, the expected person of the response (first or third person) and an interrogative versus an imperative cue in the prompt. They wrote 21 prompts in two different formats, beginning with an introductory statement and followed by either a question or an imperative. Half of the prompts had a personal manner of address and the other half had a neutral manner. The following examples are given in their paper:

Personal Question*: Some educators believe that the age at which children can legally withdraw from school should be lowered from sixteen to fourteen. What is your position on the issue of lowering the age of compulsory school attendance?*

Personal Imperative*: Some educators believe that the age at which children can legally withdraw from school should be lowered from sixteen to fourteen. Discuss your own position on the issue of lowering the age of compulsory school attendance.*

Neutral Question*: Tourism is a major industry in some states. What are the effects of tourism on a state such as Florida?*

Neutral Imperative*: Tourism is a major industry in some states. Discuss the effects of tourism on a state such as Florida.*

The prompts were administered to sophomore students in Florida colleges and universities. Participants were allowed 50 minutes to write a response to a single prompt. 929 essays were produced, with an average of 22 responses for each prompt. Each response was rated holistically and results showed that there was no statistically significant relationship between any of the prompt characteristics and the holistic scores.

In a similar study Hoetker & Brossell (1989) systematically varied the degree of rhetorical specification (length and complexity of input) in writing prompts, using brief or full rhetorical specification and personal or impersonal phrasings. An analysis of the responses revealed that there was no statistically

significant relationship between rhetorical specification and the holistic scores. The authors also concluded that full rhetorical specification had no disadvantageous effect on low-proficiency writers.

Brown, Hilgers, & Marsella (1991) also investigated how two very different types of writing prompts affected holistic scores. The main difference between this study and the ones reported above is the length of time allowed for composing, editing, and revising the written product. In total, participants were allowed 270 minutes to produce two essays in response to a reading-to-write prompt (a lengthy input text was read before test takers analyzed the text topic) and a personal experience prompt (where there was no input text). The two prompt types also incorporated different topics. The results showed that there were significant main effects for the different prompt sets. The combination of certain prompts together had a significant effect on the holistic score awarded. This is one of the few studies where holistic scores have identified significant differences in written product in response to different writing prompts. This is perhaps partly attributable to the large sample size (n=1,052) and the fact that the reading-to-write and personal prompts were substantively different in terms of the complexity and length of their input.

The studies described above are typical of ones that aimed to detect prompt effects using only holistic scores as the independent variable in a methodology that manipulates prompt characteristics as the dependent variable. Such findings suggest that holistic ratings are unsatisfactory for identifying changes in written responses that may result from differences in prompt characteristics. Findings of non-statistical significance for holistic scores may not necessarily mean that the writing does not vary at all in response to different prompts. Hamp-Lyons (1991:90) claims that "significant differences means measurably significant differences, and does not imply that the students' writing did not contain quantitative differences of text structure, rhetorical choices, lexical choices, syntactic structures, and so on." This indicates that the use of holistic scores as a measure of prompt effect may be insufficiently sensitive to capture variation in a multi-faceted skill, such as writing in a second or foreign language.

O'Loughlin & Wigglesworth (2007) offered an explanation for how holistic scoring may fail to capture variance in writing produced from differently worded prompts. After the test taker has interacted with the writing prompt the response must then be rated. It is in the rating process that prompt wording effects may be lost:

> The rater approaches the writing using either a holistic or an analytic scale or a mixture of the two. But the rater does not only interact with the students' writing; the rater also interacts with the task itself. The rater may consider the task to be more or less difficult than another task, and may compensate for this in applying the score to the writing. (O'Loughlin & Wigglesworth, 2007)

So if the interpretation of the effect of the prompt differences is made on the basis of the change in holistic score the rater may have filtered out, or completely removed the effect of the prompt. This is a view also put forward by Hamp-Lyons & Mathias (1994) and Polio & Glew (1996) and highlights a major problem with the use of holistic scores in the study of prompt effect.

The shortcomings of using only holistic scores to detect prompt effect show that an alternative approach should be adopted in this research. A triangulated approach, utilizing both quantitative and

qualitative analyses of writing samples will offer a more finely grained way of investigating the effects of differences in prompt characteristics. Additionally, the quantitative approach to the analysis of prompt effect will benefit from not relying solely on the use of holistic scores to detect variation in the written products produced in response to prompts with different characteristics, again an approach that this research will utilize.

2.1.2 Effect of prompt characteristics on analytic scores

A small number of studies have attempted an investigation of prompt effect utilizing an approach that features a main focus on the score awarded to the written responses but goes beyond a reliance on only holistic scores.

Jennings, Fox, Graves, & Shohamy (1999) used analytic scores, employing three criteria, impression of language use, content, and organization with each response being scored by two raters. Their study investigated whether a choice of topic on a writing test had an effect on the score awarded. They concluded that there was no statistically significant relationship between topic choice and writing score even though the study participants expressed a preference for being offered a choice of topic.

Smith, Hull, Land Jr., Moore, Ball, Dunham, Hickey, & Ruzich (1985) investigated the effect of three different input types of prompts on essays written by college freshmen. This study did not focus on second language writing assessment but looked in detail at how prompts that are substantively different in terms of the length and complexity of input material influence written products. Participants were given one of three different topic structures to compose an English essay within 70 minutes. Topic Structure 1 was a short, traditional independent writing prompt (p.84):

> Think about a time when you did something creative, and write an essay in which you describe that time. Then on the basis of this instance, go on to explain what a creative act seems to involve.

Topic Structure 2 required test takers to read a 206-word passage on D. H. Lawrence and a creative act. Then participants "were to describe what they thought Lawrence did and then, on the basis of what they had written, to form a generalization about what a creative act seems to involve," (p.76).

Topic Structure 3 requires test takers to read three passages by three different authors (a total of 520 words). In each passage the author describes a time when he/she did something creative. Participants were asked to describe what each of the authors did and then make a generalization about what they felt a creative act involves.

The results showed that, in some instances, the different Topic Structures produced different quality responses in terms of the ratings awarded. There were statistically significant differences between different proficiency levels of writers for all three Topic Structures. There were also statistically significant differences between Topic Structures and essay length. For Topic Structure 1, higher-proficiency writers produced significantly longer essays than weaker writers. However, Topic Structure 2 produced the shortest responses for all proficiency groups. Topic Structure 3 produced responses that

were shorter than for Topic Structure 1 but longer than for Topic Structure 2. High-proficiency writers again produced more words than other groups for Topic Structure 2. The authors concluded that Topic Structure 2 suppressed fluency for some reason.

Error counts also showed significant statistical differences between Topic Structures. Fewer errors were committed with Topic Structure 2. The authors speculated that Topic Structure 2 may give test takers some linguistic input to use in the response but it does not provide the high cognitive load that Topic Structure 3 does. The study concluded that the different Topic Structures do "make a difference in quality, fluency, and total error," (p.83).

Prompts differing in response mode (expository and narrative) have also been shown to produce significant differences in test taker writing (Quellmalz, Capell, & Chou; 1982). Participants wrote in both response modes and their responses were analytically scored using the criteria of:

(1) General impression
(2) Focus
(3) Organization
(4) Support
(5) Mechanics

The main finding was that narrative writing was harder for participants than expository writing, that is, analytic scores for narratives were significantly lower than for expository writing, especially on the general impression, focus, and organization scales. The authors go on to state that "generalizations about student writing competence must reference the particular discourse domain rather than the general domain of writing," (p.256). This study provides evidence that response mode is a prompt characteristic that may have an effect on the written product elicited, in terms of the analytic score awarded.

Weigle (1999) focused on rater behavior rather than on prompt effect, but she did report significant differences in scores awarded to two different task types. One task (graph prompt) required test takers to "interpret graphical information and make predictions based on this information," (Weigle, 1999: 150), while the other task type (argumentative prompt) asked test takers to "make and defend a choice based on information contained in a chart or table," (Weigle, 1999: 150). Hence, the two prompts differed by presentation of input material and by response mode. The responses to these prompts were scored analytically according to the criteria of content, rhetorical control, and language. Based on a Rasch analysis of the scores awarded, Weigle reported that inexperienced raters scored significantly more severely than experienced raters on the graph prompt but this was not the case for the choice prompt.

The studies described in this section show that different prompt characteristics may elicit written language that is evaluated differently based on the type of prompt. These studies investigated prompts that differed markedly in either the complexity of the input or in the discourse mode that was elicited by the prompt. It may be possible that with a sufficiently nuanced approach to analyzing the written product (analytic scores, error counts, response length), significant differences can be detected in

writing performance. Whether such findings can be replicated when analyzing relatively minor, yet systematic differences in short independent writing prompts is unclear. That will be the primary focus of the quantitative approach in this thesis.

The following section will review studies that have moved beyond the focus on variation in scores awarded to written responses. A growing body of work has now adopted a discourse analytic approach to investigating the effect of different writing prompt characteristics on written responses. This literature focuses on how different textual features of the responses may vary in response to different prompt characteristics.

2.2 Effect of prompt on discourse features of written language

Many of the studies that employed scores as the criterion measure of prompt effect have concluded that there were no significant differences in these scores, regardless of the characteristics of the prompt. However, with holistic scores failing to produce evidence of prompt effect, a different methodology has been adopted by some researchers. More recent studies have employed an analysis of discourse features of the written text in order to examine the underlying competencies required to write in a second or foreign language. These competencies include:

- Lexical sophistication – the use of original and/or low-frequency vocabulary
- Syntactic complexity – the use of lengthy and/or complex sentence-level constructions
- Accuracy – the number of errors or number of error free sentences

Examining these textual features provides a more detailed understanding of the quality of the written product than that created by the score awarded. This section will review work from the writing assessment literature that adopts a discourse analytic approach to the investigation of prompt effect.

Greenberg (1981) performed one of the most important early studies relating the prompt characteristics of cognitive demand and experiential demand to performance on a writing test. Greenberg operationalized the cognitive demand of the writing prompt as the degree of structure provided by the prompt. She used two different prompts: one high and one low cognitive demand. She operationalized experiential demand as the degree of personal experience the prompt called for. One high-level and one low-level personal demand prompt was used. The study initially employed holistic scores to assess whether differences in the prompt characteristics had any effect on writing performance. The results showed no statistically significant relationship between different prompt types and holistic scores. Despite this finding of non-significance, Greenberg also recorded a wide range of discourse measures within the essays produced in her study. She learned that:

> significant main effects were found in four of the analyses: mean number of T-units, words per clause, and words per essay. Questions with High Cognitive demands elicited more T-units, words per clause, and words per essay than did questions with Low Cognitive demands, but questions with Low Cognitive demands elicited more clauses per T-unit than did questions with High Cognitive demands. In addition, questions with High Experiential demands elicited more T-units and more words per essay than did questions with Low Experiential demands (p.72).

Although Greenberg's study suggests that overall, cognitive and experiential complexity variables will not have a significant effect on the holistic scores awarded to writing produced in a test, the discourse variables can detect variation in textual features of the response. The four prompts are presented in full in Appendix 1.

As can be seen from the prompts, there is a considerable amount of overlap between them. Indeed, it can take several careful readings to distinguish the differences between the sets of prompts. There were 192 participants in the study and each student was given 50 minutes to complete an essay. With the subtle differences in prompt types and not especially large sample size it is unsurprising that there were no statistically significant relationships between the prompt characteristics and the holistic scores. Despite the minor differences between the prompts, the discourse analyses were able to detect significant differences between textual features of the responses, especially in terms of fluency, or length of response. Greenberg's work was an early indication that a discourse analytic approach to the study of prompt effect could reveal much more about differences in written products than a reliance on differences in scores.

In a study with similar aims to that of Greenberg (1981), Hirokawa & Swales (1986) analyzed the effects of different prompt wording, requiring 32 non-native English speaking graduate students at the University of Michigan to each respond to two different writing prompts. The prompts were presented in simple and academic versions.

Simple: Would you prefer to be part of a large family or a small one?

Academic: Family size tends to vary according to a number of factors, such as culture, religion, mortality rate, and level of economic development. What are the advantages and disadvantages of small "nuclear" families as opposed to larger extended family units? State your personal preference for one of these family types and explain the reasons behind that preference.

The participants wrote one essay in 45 minutes under test conditions and then returned two weeks later to write on the other prompt under the same conditions. The two sets of responses were analyzed syntactically and the following statistically significant differences were identified (p.344). "Simple topic compositions (a) were longer than the academic topic compositions, as measured by both words written per 30 minutes and sentences written per 30 minutes; (b) contained more subordination (per standardized length); (c) exhibited greater use of the first-person, singular pronoun; and (d) contained more morphological errors."

Zhang (1987) investigated whether six different prompts, designed to be at different levels of cognitive complexity elicited written language that varied according to three criteria; fluency (number of words), syntactic complexity (average sentence length and clauses per sentence), and accuracy (errors per 100 words). The six prompts were administered to 63 ESL learners at The University of Hawaii and a small college, also in Hawaii. Zhang operationalized cognitive complexity in two ways; first, questions that required only a factual response versus ones that were interpretive, "requiring an extrapolation from the immediate factual content," (Zhang, 1987: 472), and second, what Zhang terms convergent recall

questions versus divergent, above-recall questions. The explanations provided in the paper do not make it particularly easy to differentiate between the two categories of cognitive complexity and there appears to be some overlap between the categories. Zhang (1987; 477) claimed that:

- More cognitively demanding questions elicit significantly longer responses.
- More cognitively demanding questions elicit responses with significantly longer sentences and more cases of syntactic coordination and subordination.
- There are no significant differences in the accuracy of the responses to the questions at different levels of cognitive complexity.

The difficulties with the operationalization of cognitive complexity in the research design make it wise to be conservative in the extent to which the findings of Zhang's study can be generalized; however, this paper provides another indication that fluency and syntactic complexity are textual features that may vary depending on prompt characteristics. As accuracy is such an important aspect of second language writing proficiency, it is difficult to consider omitting it from the analysis of written product, even though Zhang's work does not report any significant differences in accuracy.

Spaan (1990) worked within the same assessment context (the MELAB) and with a similar population to that of this current research. Spaan investigated how two prompt characteristics (narrative/personal or argumentative/impersonal) affected both score awarded and a range of discourse measures within the written responses. The prompt characteristics of person of response and response mode will be shown to be key prompt characteristics in Chapter 3. Four writing prompts in two paired sets were analyzed and 176 essays were collected.

Spaan reported findings for the whole sample and several subgroups within the sample population but the sample sizes for the subgroups are very small (in the 20s and 30s) so only whole sample findings will be reviewed here. The difference in mean scores between the prompt types was less than one point out of the 97-point scale, indicating that there was virtually no effect of the prompt characteristics on score awarded. Spaan also applied a range of discourse measures to the written responses. The discourse measures were selected to operationalize the textual features of fluency, syntactic sophistication and accuracy, and lexical range and sophistication. In addition, a rating scale was adopted from Connor and Lauer (1988) to measure rhetoric. However, these discourse measures were only applied to a subgroup of the population with a sample size of 21 test takers, making the findings of little interest. The current research will investigate similar prompt characteristics to those of interest to Spaan and will employ similar discourse measures; however, this work will involve a larger sample size overall and in response to each prompt. The larger sample size will hopefully make the findings of this study more generalizable and also allow for a variety of statistical analyses to be meaningfully applied to the dataset.

Way, Joiner, & Seaman (2000) performed a study of three different writing prompts with learners of French in California public schools. The writing prompts were described as a bare prompt, a vocabulary prompt, and a prose prompt. The bare prompt was a very short independent writing prompt, presented to the test takers in English requesting a letter be written in French to a pen pal. The vocabulary prompt contained the same task as the bare prompt but the vocabulary prompt provided participants with a list of French vocabulary and English definitions. The prose prompt required the participants to read a letter

from a pen pal, written in French and then write a response in French. There were also three different writing tasks within the three different prompt types; a descriptive task, a narrative task, and an expository task. Three different responses were collected from each participant and 937 samples were analyzed by holistic score, length of product, mean length of T-units, and percentage of correct T-units. These discourse measures operationalized the constructs of fluency, syntactic complexity, and accuracy.

There were numerous statistically significant differences observed in the responses between the different prompt types and different tasks. The descriptive writing task elicited significantly higher scoring responses than the narrative and expository tasks. The expository task responses were scored significantly lower than the responses to the other tasks. The responses to the bare prompts were scored significantly lower than the responses to the other prompts. Descriptive responses were the longest and the prose prompt elicited statistically significantly longer responses than the bare prompt. The prose prompt also elicited significantly more syntactically complex responses than the other two prompt types. The bare prompt elicited the least syntactically complex responses and these were significantly less complex than the responses to the vocabulary prompts. Responses to the prose prompt were significantly more accurate than the responses to the other tasks. Descriptive and narrative responses were both significantly more accurate than expository responses.

This study reveals a large number of statistically significant differences in written product, in response to a range of prompts of different complexity and response modes. Although the assessment context is different from that of other studies (French language in the US classroom), the findings suggest that prompts of different complexity and response mode may elicit significantly different responses. The textual features of second language writing proficiency that were operationalized in Way, Joiner, & Seaman's study (fluency, syntactic complexity, and accuracy) will also be applied to the essays in this study as they are strong predictors of overall second language writing proficiency. Chapter 4 will describe the selection of the discourse measures and the characteristics of the prompt that are investigated in this work.

Although the study by Cumming, Kantor, Baba, Erdosy, Keanre, & James (2005) looked at integrated writing tasks (on the TOEFL) it is of great relevance to this study because of the highly detailed approach it took to the selection of discourse measures. The level of detail seen in the operationalization of second language writing proficiency provided a step forward for the field of assessing writing.

Integrated writing tasks require the test taker to read or listen to some input material (an academic text or short lecture for the TOEFL) and compose a response to what was heard or read. Independent writing tasks require test takers to respond to a short written prompt. Participants were required to write six essays each, in response to both independent and integrated tasks. These responses were coded in detail for:

- Lexical and syntactic complexity
- Grammatical accuracy
- Argument structure
- Orientation to evidence
- Verbatim use of source text

Participants were sorted into three separate proficiency levels and significant differences were found between responses to the independent and integrated tasks. Significant differences in textual features were identified in lexical complexity (text length, word length, ratio of different words to total words written), syntactic complexity (number of words per T-unit, number of clauses per T-unit), rhetoric (quality of propositions, claims, data, warrants, and oppositions in argument structure), and pragmatics (orientations to source evidence in respect to self or others and to phrasing the message as either declarations, paraphrases, or summaries), (p.5).

Cumming et al.'s work stands out for the detailed discourse measures that were applied to the written products. This approach revealed numerous significant differences in the responses to the varied integrated tasks. The focus on integrated tasks reduces the relevance of Cumming et al.'s work to this thesis but the broad range of discourse measures applied is very informative.

O'Loughlin & Wigglesworth (2007) examined five different versions of two experimental tasks of varying cognitive complexity (one containing 16 pieces of information and the other with 32 pieces of information). 210 ESL students completed four different tasks (two each at the different levels of complexity). They were given 20 minutes to complete each writing task. The participants were sorted by proficiency and the responses were double rated using both a global band score (a holistic rating) and an analytic scale. The analytic ratings were performed according to three main criteria; task fulfillment, coherence and cohesion, and vocabulary and sentence structure. The main conclusion drawn was that, "the results of these quantitative analyses reveal that the differences elicited by the different amounts of information provided in these tasks, and the different types of presentation are very small," (p.390).

This finding prompted the researchers to analyze the discourse within the responses to see whether there was any systematic difference in written performance when evaluated using a range of discourse measures. The authors looked at task fulfillment (number of words and accuracy of information), coherence and cohesion (structure and organization of the body, conjunctive and referential cohesion), and vocabulary and sentence structure (number of clauses, type of clauses, number of T-units, number of error free clauses and T-units, and repetition of key words). The discourse analyses revealed that the task with less input (16 pieces of information) produced responses with greater complexity on most measures (structure, organization, cohesion, subordination, and repetition of key words) across all proficiency levels. As with Cumming et al. (2005), the work of O'Loughlin & Wigglesworth confirmed the merit of applying a broad range of discourse measures in order to create a layered view of the written product being analyzed.

Kuiken & Vedder (2008) investigated how prompts of differing cognitive complexity affected a range of discourse measures. They administered two different prompts to 91 Dutch learners of Italian and 76 Dutch learners of French. The prompts were designed under a "complex condition" and a "non-complex condition" (Kuiken & Vedder, 2008: 52). Under the complex condition, the writers had to fulfill several more requirements to complete the task than under the non-complex condition. The responses were analyzed for accuracy (total number of errors and error severity), syntactic complexity, and lexical variation. The results indicated that fewer mistakes were made by both groups of language learners (French and Italian) in response to the complex condition than in response to the non-complex

condition. There were significantly fewer total errors made and also significantly fewer severe errors in response to the complex condition. However, there were no significant differences in either syntactic complexity or lexical variation for either language group. That is, the written language elicited by the two prompt types was stable as measured by the syntactic complexity and lexical variation measures. The overall conclusion drawn by the authors was that increasing task complexity may elicit written language that is increasingly accurate for second language writers under exam conditions.

Discourse analytic studies that have addressed prompt effect have drawn differing conclusions when investigating accuracy. While Kuiken & Vedder (2008), along with Way, Joiner, & Seaman (2000) observed variation in the accuracy of responses between different prompts, other studies (Zhang, 1987) found no significant differences in accuracy. The accuracy of second language writing is a key feature of the quality of the written product. Although prompt effect studies that have investigated the accuracy of responses are inconsistent in terms of the findings they have produced, the importance of accuracy to the construct or writing proficiency in a second or foreign language is so great that accuracy is a key writing competency that will be investigated in this study.

Ong & Zhang (2010) administered writing prompts of varying task complexity to Chinese learners of English in Singapore. Task complexity was operationalized quite differently to the approach of other studies. The length of planning time before composing was varied, with three independent groups being allowed 20 minutes, 10 minutes or no planning time before beginning to write. The level of detail provided within the prompt was also divided into three distinct conditions, "(1) topic, ideas, and macro-structure given; (2) topic and ideas given; and (3) topic given", (Ong & Zhang, 2010: 223). The final task complexity variable that was manipulated was the condition under which participants could revise their first drafts before submitting a completed piece of writing. There were only two conditions; revising with first draft visible and creating a final draft without being able to refer to the first draft. All three prompt variables were intended to provide differing levels of cognitive complexity. The responses were analyzed for fluency (total number of words per minute of writing and the mean number of words produced for each minute of total time spent on the task). Lexical complexity was operationalized as type-token ratio adjusted for text length.

Ong & Zhang found no significant differences for the first measure of fluency (total number of words per minute) for any task conditions. There was a significant difference in the second fluency measure (mean number of words per minute for total task time) based on the planning time task condition. Participants who did not have planning time produced significantly more words per minute on task than participants under the other planning conditions. Planning time also had a significant effect on lexical complexity. Again, responses produced without planning time were significantly more lexically complex than responses produced with either ten or twenty minutes planning time. Both of these findings are somewhat surprising as one may expect planning time to allow second language writers to prepare a lengthier response and to marshal their linguistic resources with less time pressure and hence make full use of their lexical resources. This was not the case in Ong & Zhang's study and the researchers speculated that the participants may not have utilized planning time effectively.

Finally, Lim (2010) utilized a different approach to studying prompt effect to the work reported above. Lim developed a detailed prompt categorization framework and then used the Rasch model to create estimates of prompt difficulty based on the holistic scores awarded to MELAB essays. The difficulty of 60 distinct writing prompts that varied according to five key prompt characteristics (see Chapter 3 for

details of the categories) were analyzed. Lim reported that the null hypothesis of all prompts being of equal difficulty was rejected, indicating that the prompts were not consistently equivalent, with a range of 2.78 logits between the easiest and the most difficult of the prompts (Lim, 2010: 105). Lim suggested that 14 of the 60 prompts analyzed should be removed from the operational prompt pool in order to establish a narrower prompt difficulty range (1.47 logits) and to better establish a claim of prompt equivalence. This provides some indication that there was variation in the equivalence of MELAB writing prompts analyzed that could potentially have contributed to differential test performance in terms of the score awarded.

In terms of the prompt characteristics in Lim's categorization framework, the characteristic that exhibited most difference was topic domain. The largest differences were between prompts in the business and social domains). Lim concluded that scores awarded to the different prompts were stable apart from a few outlying prompts, with only topic domain having a potentially significant effect on scores. However, a careful reading of the results suggests that the detailed prompt categorization framework devised by Lim holds great promise for learning about prompt characteristics that contribute to prompt equivalence, or the lack thereof. Coupled with a discourse analytic approach to the analysis of written responses, the detailed categorization of writing prompts provides a promising way of taking a nuanced look at prompt equivalence and such an approach (detailed prompt categorization and discourse analyses) will be adopted in this research. A literature review of prompt categorization frameworks will be provided in Chapter 3 and an account of how writing prompts are categorized in this thesis will be given in Chapter 4.

2.3 Summary of literature on effect of prompt characteristics on written product

The literature reviewed in the sections above makes clear that different prompt characteristics can indeed have an effect on written product. Many of the early studies of prompt effect focused on how different prompt characteristics affected holistic scores awarded to written responses. This was not a fruitful approach as these studies commonly found that there were no statistically significant differences in holistic scores, suggesting that there was no identifiable effect of the prompt characteristics.

Studies that went beyond the use of holistic scores came to different conclusions, revealing that a variety of prompt characteristics can have measurable effects on the written product. The most straightforward finding that these studies have in common is that prompts of differing levels of cognitive complexity, particularly in terms of length of input, may elicit responses that vary by some key textual features of written language. In particular, studies that administered independent or relatively bare writing prompts alongside integrated, or reading-to-write prompts to large groups of language learners found that the responses to these different prompts varied by a range of discourse measures, such as those that operationalize the traits of lexical sophistication, syntactic complexity, fluency, and accuracy.

The literature on prompt effect is not extensive, and the number of studies utilizing a discourse analytic approach remains limited, but the findings suggest that relatively substantive differences in the complexity of writing prompts will elicit written language that may vary by a range of textual features. However, a number of issues remain unresolved. While independent and integrated writing prompts

may elicit written language that differs by a range of textual features there is, as yet little evidence that independent writing prompts that differ by certain prompt characteristics will exhibit similar levels of prompt effect such as that found between independent and integrated prompts.

There is little evidence in the literature that different independent prompt characteristics elicit written language that varies systematically. Also, the conclusions that may be drawn about the specific prompt characteristics that contribute significantly to prompt effect are very limited. Length of input, when comparing reading-to-write prompts with independent prompts appears to be one generalizable feature of prompts that contributes to prompt effect but there is little consensus in the work reported above regarding any other characteristics that systematically contribute to prompt effect.

The specific relationship between common characteristics of independent writing prompts and textual features of the written product is not yet well understood. There is no consistency in the literature described above that shows designers of writing prompts how varying common prompt characteristics will influence the written language elicited. This lack of clarity in the literature makes the establishing of prompt equivalence in writing tests elusive. The field needs to develop a systematic understanding of the relationship between prompt characteristics and textual features of the written product. The writing assessment field lacks a common basis for capturing the effects of the prompt on written products and this study aims to make a positive contribution toward a systematic approach to the study of prompt effect.

One feature all of the studies described above have in common is that they utilized quantitative methods to investigate prompt effect. The quantitative findings are valuable because the written product is ultimately what will be read and assessed in either a classroom or standardized testing context. However, a fuller understanding of the effect of writing prompts may be come to by bringing the writer more fully into the picture. The written product is created during an interaction between the writer and the prompt. An analysis of the final written product does not capture the processes that were undertaken by the writer. For example these processes may include choosing a prompt to respond to, planning and organizing a response, composing a response, and editing the response. A purely quantitative approach to investigating prompt effect will ignore these human interactions with the prompt. While the analysis of the final written product may indeed be the priority to understand prompt effect, ignoring the human interactions with the prompt seems unwise.

A better understanding of what the writer or test taker does with the prompt in the act of creating the written product can only help researchers to gain more insight into the relationships between prompt, writer, and product. This research will attempt to investigate this relationship more deeply than the quantitative studies reported above by adopting a mixed-methods approach; a quantitative approach that builds on the literature reported in this chapter and a qualitative approach that will focus on the test taker (the writer) and aims to explore how the individual interacts with the writing prompt and which prompt characteristics are important to the individual when creating the written product.

The current work will build on the literature reviewed above, concentrating in four main areas:

- A focus on independent writing prompts and an understanding of the characteristics that contribute to prompt equivalence.
- An understanding of how variation in key prompt characteristics may have an effect on the written product.
- A systematic approach to prompt categorization, based on the literature (see Chapter 3). When effects are detected, they can be traced to prompt characteristics that are common across many prompts.
- A mixed methods approach that retains the best of the quantitative approaches described above, but also utilizes a qualitative approach capable of taking into account the reactions of writers and test takers to different writing prompts.

The following section will briefly review the literature on other factors that may affect how test takers perform on a writing assessment. These factors need to be understood and controlled for when investigating prompt effect as they may potentially be confounding variables in any analysis of writing proficiency, especially the second language writing context that is the focus of this work. This literature does not directly explain the interaction between test taker and writing prompt but it does help explain a range of factors that may interfere with the prompt/test taker relationship that is at the heart of this thesis.

2.4 Test taker factors

The literature on test taker factors that can affect performance on writing tests (see Barkaoui, 2007 for a meta-analysis of relevant literature) suggests several different test taker factors that should be considered. The factors most commonly agreed on are language proficiency (Ruth & Murphy, 1984; Hoetker, 1992; Tavakoli, 2009), world knowledge (Ruth & Murphy, 1988; Read, 1990; Tavakoli, 2009); cultural and linguistic background (Ruth & Murphy, 1984; Hamp-Lyons, 1988; Johns, 1991; Kroll & Reid, 1994; He & Shi, 2008; Tavakoli, 2009) and educational background, including exposure to writing training (Ruth & Murphy, 1988; Johns, 1991).

2.4.1 Test taker factors – language proficiency

The language proficiency of the test taker can interact with prompt characteristics to have an effect on writing performance. The first interaction with a writing prompt is the act of reading and the reading proficiency of the test taker will partially determine the understanding of the prompt that is constructed by test takers. The more linguistically complex the language of the prompt, the more disadvantaged the weak L2 readers will be. Disparities in reading proficiency may cause misinterpretations of the prompt and create unexpected responses that may not be based on the assumptions of the prompt designer or the raters (Kroll & Reid, 1994). Misinterpretations of the prompt may lead to potential construct irrelevant variance (the effect of reading proficiency on writing test performance) and pose a threat to the validity of writing tests, especially for lower proficiency candidates.

The degree of rhetorical specification (linguistic and cognitive complexity) in the prompt and the intended mode of response are two aspects of writing assessment that may interact with the proficiency level of test takers. The amount of rhetorical context that should be provided in writing prompts is a matter of debate. While there are arguments in support of full rhetorical specification, typically called

for to provide the test taker with an authentic context to write within, the amount of language that must be processed in this case may be overwhelming for lower proficiency test takers (Hoetker, 1982: 386). Ruth & Murphy (1984:419) argued that as writers mature and develop they may go beyond the rhetorical demands set by the prompt and want to demonstrate the wider discourse skills they have control of. Weaker test takers may be unable to accurately process a prompt with extensive rhetorical specification or a lengthy input text, possibly leading them to respond in a way not expected by the test maker. Conversely, high-proficiency test takers may wish to demonstrate their abilities by going beyond what the prompt asks for.

Another related issue is that student understanding of the intentions of the prompt may also override the expected mode (argumentative, expository or narrative) of the response (Hoetker, 1982). Hoetker claimed that the mode of writing any essay topic calls for is "precisely that mode that any particular student interprets it as calling for" (Hoetker, 1982: 379). This interpretation will be affected by the proficiency of the individual test taker. Response errors in terms of the specified mode may be due to misinterpretation of the prompt by weak readers, an inability to write in a more demanding mode (typically argumentative writing is seen as more difficult), or the desire of stronger test takers to write in a different mode to that specified in the prompt. Hoetker believed that some writers may not be capable or willing to write in a particular mode even if that is called for by the prompt. Hoetker's work was with first language, not ESL learners, so the extent to which his findings are generalizable to second language learners is a matter of conjecture. However, it seems possible that second language writers may be more likely to be susceptible to response errors as a result of misinterpreting the prompt than native speakers, either due to weak language resources in their L2 or because they are unfamiliar with the particular conventions of a specific mode of writing in a different language context.

The cognitive complexity of writing prompts and the response mode are two prompt characteristics that will be addressed in this work. As such, the research design must take into account and control for participant language proficiency. If the language proficiency of the sample population is not controlled for, the ability to interpret the relationships between the prompts and the test takers may be jeopardized. For example, if the test takers who respond to one prompt are all high-proficiency language learners and the test takers who respond to another prompt are all low-proficiency language learners, the findings of any analysis of written product of responses from quite different populations will be attributable to the differences of the populations as much as any differences in the prompts.

2.4.2 Test taker factors – world knowledge and linguistic background

Ruth & Murphy (1984: 413) claimed that the meaning of any particular writing prompt depends on the "linguistic, cognitive, and social reverberations set off in the respondents. Both the language of the topic and the general knowledge of the participants interact in a writing test to determine what meanings the topic may elicit." The intended meaning of the prompt may not be the same as the understood meaning of the test taker. The understood meaning will be dependent upon a range of test taker factors such as, "inadequate control of linguistic and semantic knowledge, weak commitment to succeeding on the test, inadequate world knowledge, and inexperience with testing contexts and conventions," (Ruth & Murphy, 1984: 415). Read (1990: 110) believed similarly that the prompt topic "should be about a

subject that all potential test takers have enough relevant information on, or opinions of, to be able to write to the best of their ability. On the other hand the topic should not be too simple or predictable." Early research in the field of topic effect suggested that familiarity with an essay topic advantaged those candidates and the depth of world knowledge a test taker brings is likely to be relevant to the sample of writing produced.

Test taker interviews performed with GRE candidates (Powers & Fowles, 1998) suggested that topic interest and topic knowledge were important variables in how test takers construct the difficulty of writing prompts. The test takers interviewed (Hispanic or African American test takers) consistently reported that they preferred prompts they could identify with, that drew on personal experience, that were clearly stated, that elicited strong feelings, and that were interesting (Powers & Fowles, 1988: 9). The sampled group described having most difficulty responding to prompts where they felt they lacked familiarity, knowledge or appropriate background along with prompts that were unclear or ambiguous, (p.11).

Research into the cultural and educational experiences of IELTS test takers indicated that Chinese and Greek students performed differently on the exam. Mayor, Hewings, North, Swann, & Coffin (2007) found that "low-scoring Chinese L1 candidates made significantly more grammatical errors than Greek L1 at the same level of performance," (p.251). The study reported relatively minor differences in argument structure between the two groups but reported that "Chinese L1 writers also have a greater tendency to directly address the reader and to speak in the collective voice, (p.300).

ETS research reports on the TOEFL (see, for example, Breland, Lee, Najarian, & Muraki; 2004; Lee, Breland, & Muraki; 2004) consistently concluded that there are only very small statistically significant differences in difficulty across TOEFL writing prompts. TOEFL Research Report 76 (Breland, Lee, Najarian, & Muraki; 2004), which looked at writing prompt difficulty by gender found that there were statistically significant differences between male and female candidate performance on almost all writing prompts, but that the effect sizes were sufficiently small to be of little or no concern. The effect sizes were generally around 0.2 standard deviations of the overall score. While this does seem low, the fact that female candidates consistently scored better than expected and male candidates worse than expected, after controlling for language proficiency suggests that this may be an aspect of prompt design that is of potential concern and to be considered and screened for during pretesting. It also suggest that sample populations of studies into writing prompts should be balanced in terms of the gender of the population.

TOEFL Research Report 77 (Lee, Breland, & Muraki; 2004) examined the comparability of TOEFL writing prompts for different first language groups. It focused on Asian language speakers versus European language speakers because of their differing performance overall on the TOEFL. The report concluded that there were no significant group effects between the two in the writing section after adjusting for overall language proficiency. However, in 27 of the 81 prompts analyzed there was "a statistically significant amount of variation in essay scores," (p.14) as a result of native language effect. The study stated that the effect sizes were negligible to overall group performance but the fact that around one

third of TOEFL writing prompts resulted in first language background interfering with test takers' true scores is troubling.

The studies reported in this section show that the makeup of the sample population is important when investigating complex relationships between test takers, writing prompts, and written products. The greater diversity there is in the sample population, in terms of linguistic background and world knowledge, the less risk there will be that the results are influenced by confounding variables from the sample population. However, even diversity within the sample population may not completely guard against individual effects.

2.4.3 Test taker factors – cultural background

Test takers will not all necessarily come to an exam with the same cultural and linguistic frame of reference (Ruth & Murphy, 1984). This is especially true of international English language proficiency tests, which are taken by candidates from many different countries. These differing cultural backgrounds can potentially cause different interpretations of writing prompts and lead to unexpected responses. An example given by Ruth and Murphy is that of a writing test that asks candidates to produce a "friendly letter." Test takers who are unfamiliar with the expected conventions of the formal testing situation may produce a letter that is too friendly and hence be penalized for a misunderstanding of standardized testing conditions.

A similar point is made by Kroll & Reid (1994: 236), who stated that cultural interference can cause test takers to misconstrue or even miss the point of writing prompts. Those without the assumed cultural reference could be disadvantaged by certain writing prompts; especially those which assume a knowledge of western cultural values. The authors cited the example of a prompt that asked test takers to respond to a topic about a "blind date." Several ESL candidates interpreted this literally as asking about a date who could not see, leading to a number of unexpected responses.

Connor & Kramer (1995) reported that Korean, Belgian, Bolivian, and American students in a US business school interpreted and responded in different ways to writing assignments set in their program. Connor and Kramer concluded that the ESL students differed from native English speaking students in task representation. Task representation refers to the process of interpreting the writing prompt and then using it to organize a response. Interviews with the participants suggested that the reasons for the differing task representations were both cultural and educational. The Korean participant differed most markedly from the task expectations and his difficulties were seen as typifying those faced by Asian language speakers when presented with academic writing tasks.

Johns (1991) explored in great detail the difficulties of an ESL student attempting to pass a standardized writing test. The Asian test taker reported having difficulty interpreting the instructions of the writing test. He also felt uncomfortable about the need for argumentation in the written response to the prompt. As a science student he was familiar with the demands of his field but was unsure of how to frame an opinion without scientific backing for an English writing assignment. These are clear examples of the types of pragmatic competence that Hamp-Lyons (1988) described as being potentially lacking in test takers unfamiliar with western academic traditions. The subject in Johns' study described a lack of

knowledge of certain topics that appeared on the writing assessment he experienced. In these instances he reported being unable to focus on skills such as organization, as his linguistic resources were focused on overcoming his knowledge gap with the prompt topic. This test taker clearly felt he was disadvantaged by some topics and his lack of background knowledge and awareness of academic expectations contributed to his poor test performance.

A more recent study found somewhat similar conclusions to Johns' work. He & Shi (2008) interviewed 16 international students (from mainland China and Taiwan) who complained about culturally biased essay prompts and a lack of understanding of the requirements of the essay test. Several test takers described topics on plastic surgery, teenage crime, and divorce rates as requiring knowledge of their country of residence, Canada, which they felt they had little cultural knowledge of and as a result, little relevant language to respond with.

The findings of the literature reported in this section are based on qualitative methods; primarily interviews with test takers in the case of Johns (1991), Connor & Kramer (1995), and He & Shi (2008). The studies focus on how test takers engage with a writing assessment and detail the difficulties they face with a focus on cultural background and resulting lack of familiarity with standardized writing tests. These studies, although based on a very small number of test takers indicate that some individuals may respond to writing prompts in ways that are unforeseen by test designers. The ways that test takers engage with writing prompts and how they go about responding to them is an under researched area in second language writing assessment. The dynamic of the test environment and the interaction between the prompt characteristics and the test taker is a potentially rich source for gaining a better understanding of how different writing prompts may have an effect on writing test performance and will be a key part of this thesis.

2.4.4 Other test taker factors

Age, gender, and ethnicity have also been reported as factors that can influence scores on writing tests. Gabrielson, Gordon, & Engelhard Jr. (1995) explored how a choice of writing prompts influenced performance on a writing test for 11-th grade students in an American high school. 20 writing tasks were assigned to a total of 34,200 students in Georgia with the population being 52% female, 48% male, 67% white, and 33% black. The students were asked to complete a persuasive piece of writing within 90 minutes. Responses were analytically scored using four criteria: content and organization, style, conventions, and sentence formation. The main finding was that "the effects of the student characteristics on the essay scores are much greater than the effects of writing task," (p.281).

The literature reviewed above (Sections 2.4.1-2.4.4) indicates that several test taker factors (linguistic, cultural, and educational background, along with other factors such as age, gender, and ethnicity) may have the potential to affect individual performance on a writing test. These factors will be considered when collecting written responses to be analyzed in this research. Efforts will be made to ensure that the sample population the responses are drawn from will be diverse in language background, nationality, gender, and age. A sample population that is diverse in these factors will help minimize the confounding variables that may interfere with understanding the complex relationships between test

takers, prompts, and written products. The sample population that participated in this study is described in Section 4.4.1.

2.5 Examination context variables

The final set of variables that can affect the responses produced in a test of writing proficiency are examination context variables. These variables include the time allowed for writing, comfort and environment of the test location, test taker motivation, and transcription mode (paper and pencil or computer). If the test situation is to be fair to test takers, they ideally need a quiet and comfortable place to write, a reasonable length of time to respond to the prompt, and a way of writing a response (paper and pencil or a computer) with which they are comfortable.

2.5.1 Time constraints

Most ESL writing proficiency tests limit the amount of time test takers have to produce a written response. Connor & Carrell (1993: 143) suggested that test takers must "make decisions about whether they are to address a specific audience, what the purpose of the writing is, what style and tone is expected, what length is expected" all within the time limit set for the exam. Ruth & Murphy (1988) claimed that test time limits interact with candidates' writing proficiency to influence written output. They described high proficiency writers typically taking a long time to plan their responses and then consequently feeling rushed when they begin writing. Ruth & Murphy (p.152) suggested that this time pressure may prevent the more proficient writers from having time to re-read or proofread their own work. Test takers had 30 minutes for the whole writing test (including reading, planning, and composing) in Ruth & Murphy's experiment. They concluded that although more proficient writers were still likely to score more highly on speeded writing tests, such a testing context is "likely to truncate severely the range of performance elicited during the assessment, (p.153).

Hall (1991) reported in Barkaoui (2007) claimed that texts produced by writers under test conditions, including a time limit could differ substantially from those that the same writer could compose under non-test conditions. Barkaoui (2007: 117) highlighted the effects of time pressure on test-taking processes as one aspect of second language writing assessment that requires further research.

ETS performed research into the effect of time constraints on Test of Written English scores (Hale, 1992) and found that increasing the time allowed for writing from 30 to 45 minutes resulted in a modest but consistent increase in scores. A student survey also showed that test takers thought 45 minutes was a more appropriate length of time for the test than 30 minutes, with low-proficiency learners being more likely to report 30 minutes as insufficient.

2.5.2 Test taker motivation

Test takers are most likely to perform to the best of their abilities when the test result is important to them. Candidates who do not wish to take the test but are taking it for some extrinsic reason (pressure from teachers or parents) may not perform to their true ability level (Ruth & Murphy, 1984: 415).

However, test takers who feel severely pressured to pass or excel at a particular test may suffer from test anxiety that can prevent them from showing their true ability.

Another aspect of motivation to write well is the level of interest generated by the prompt. Evans (1979) reported in Ruth & Murphy (1988) described some essay topics failing to generate any interest or enthusiasm among test takers. This is a theme taken up by Petersen (2009) who investigated test takers who deviate from the given prompt because they may find the constraints of the timed essay test to be restrictive and the experience of taking such a test to be "dreary and formulaic, (p.192). Peyton et al. (1990: 159) made a related point when they claimed that "ESL students' demonstration of linguistic complexity in writing may be enhanced at least in part by opportunities to communicate real messages, about topics they are familiar with, to an audience they know." Prompts that require test takers to write with no particular communicative purpose and to no particular audience may not "generate the levels of linguistic complexity that the students were capable of", (p.159).

2.5.3 Transcription mode

Whithaus, Harrison, & Midyette (2008) looked at the influence of keyboarding versus handwriting in a high-stakes writing assessment. Students in their study were presented with a choice of taking a writing exam by computer or handwriting. Students and raters were surveyed for their opinions concerning the two different approaches. Their findings indicate that:

> (1) students see the medium of composing as influencing the quality of their writing;
>
> (2) raters' attitudes toward reading handwritten or keyboarded essays can be complex, contradictory, and sometimes at odds with what the administrative data reveal about program-level trends; and
>
> (3) students and raters have different perceptions of how writing quality varies and errors occur in handwritten and keyboarded essays

Apart from the first month the students were offered a choice of transcription modes, a majority of the test takers selected to write using a keyboard. Around 40% of the test takers still chose to write by hand in each test administration though. The reasons cited by students for preferring keyboarding were that it was faster, more familiar and legible, and that errors were more recognizable. "Overall, they perceived the presentation of their work as cleaner in the computerized version because they could edit and proofread more effectively on the screen," (p. 16). The raters held opposing views about the handwritten versus keyboarded texts. The raters believed that the keyboarded responses contained a greater number of errors (particularly surface errors such as spelling and mechanics) than the handwritten ones. The authors believed that raters may have higher expectations for keyboarded responses than for handwritten ones in that they expect keyboarded writing to have gone beyond the first draft stage and be more accurate because it has been written with a computer.

2.6 Summary

The literature reviewed in this chapter makes clear that the sample of writing elicited by writing prompts is affected by a wide range of factors other than just the influence of the prompt. These factors need to be controlled, to the greatest extent possible in order for the effect of the prompt on the written product to be investigated without confounding variables interfering with the effect of the prompt. Firstly, there are test taker factors, which may interact with the writing prompt and affect the response produced. The language proficiency, world knowledge, and cultural and linguistic background of the test taker may all contribute to the final text produced. For researchers to make informed judgments about how writing prompts affect written responses, it will be valuable to have as much information as possible about the sample population they are studying and for the sample population to be as diverse as possible.

Examination context variables should not be discounted by researchers. The time allowed for a writing test would seem to play an important role in the ability of test takers to fully demonstrate their writing proficiency. The test environment and the motivation of participants in a study of prompt effect will ideally be as close as possible to that of a high-stakes writing assessment. If the participants are not motivated to fully demonstrate their true level of writing proficiency, the meaningfulness of the findings will be at risk.

From the review of literature in this chapter, it is clear there are many factors and variables for the writing assessment researcher to consider when designing a study of prompt effect. A triangulated approach to investigating different writing prompts, utilizing a mixed methods approach of quantitative and qualitative techniques, while controlling for test taker and test context variables has much to offer the field of second language writing assessment. A discourse analytic approach to detecting significant differences in written product has proven the most effective way of identifying differences in key linguistic features of written language. This quantitative approach, coupled with a qualitative approach that explores how individual test takers interact with and respond to different prompts will take the best of the existing literature on prompt effect. These are the methods and aims of the current research based on the review of literature presented in this chapter.

Chapter 3 – Review of Literature (2)

In Chapter 2, I reviewed the literature on prompt effects. This revealed that some aspects of writing prompts appear to affect features of the responses they elicit. However, it was not clear which types of prompt were most likely to contribute to the effects observed. In this chapter, I review the literature on the categorization of writing prompts, with a view to better understanding which characteristics can most consistently distinguish between different writing prompts.

I will review relevant literature from three different perspectives.

1. Task complexity frameworks, which provide an understanding in a broader assessment context of the features of productive language tasks that may have an effect on the language elicited.
2. Studies that have explicitly studied writing prompt categorization as their primary focus.
3. Studies that have incorporated a categorization of writing prompts as a secondary focus.

Following a discussion of what can be generalized from the literature, I will draw conclusions regarding the identification of key prompt characteristics.

One of the overarching goals of this thesis is to better understand how prompt equivalence can be established within a writing test. Establishing prompt equivalence is important so test takers are consistently presented with writing prompts that provide them an equivalent opportunity to demonstrate their writing proficiency. Establishing prompt equivalence is important for test fairness and for score interpretation. In order to achieve prompt equivalence it will be helpful to have some insight into which aspects of the writing prompt may contribute to differential test performance. Differing test performance may manifest itself via varying test scores and if the same writer may be awarded different scores depending on which prompt he or she responds to, the prompt is introducing measurement error to the process of evaluating writing proficiency. This outcome would bring into question the validity of the test scores as the score would depend, to some extent on the prompt and not on the ability of the test taker.

Differing test performance may also manifest itself via writers producing responses to prompts that vary significantly in key textual features. If certain prompt characteristics elicit responses that are significantly different from responses to other prompts with distinctly different characteristics, these systematic differences may create difficulties for score interpretation. Although these differing responses may still result in the same or very similar scores, the scores may have been arrived at for quite different reasons.

The score awarded to a piece of writing is interpreted as a representation of the test taker's writing abilities. But what abilities are represented in that score? The score may mean that a test taker is capable of writing grammatically accurate but short and simple texts. But the same score could mean that a test taker may be able to write a lengthy text, using some sophisticated vocabulary but that is challenging to process for meaning because it lacks control over syntax. Without a concrete understanding of the abilities that a score represents, that score is potentially of limited value to both the test taker and the test stakeholders, such as the institutions that make decisions about test takers

based on the scores. Whether the writing prompt influences the score awarded and/or the stability of the textual features of the response, test performance is dependent to a greater or lesser degree on the prompt. If writing performance varies significantly and systematically based on the prompt, this is potentially a source of measurement error, which is undesirable in any form of assessment and especially for high-stakes assessments.

One of the first steps toward establishing prompt equivalence is identifying the characteristics of the prompt that commonly distinguish one prompt from another. The common characteristics that differentiate between prompts are likely to be those that need to be controlled if prompt equivalence is to be established and maintained. Gaining an understanding of how the existing literature has approached the categorization of writing prompts will help to identify the key characteristics of writing prompts, or at least shed light on how other researchers have attempted to categorize writing prompts. This is the literature that will be reviewed in this chapter.

Within the context of this study, a clear understanding of how writing prompts may be categorized is important. A prompt categorization framework provides insights into the distinguishing characteristics of different writing prompts and these characteristics, the ones that distinguish one prompt from another will aid in identifying prompts that may most contribute to prompt effect. Chapter 4 will outline how the prompts investigated in this study were selected (see Section 4.3).

3.1 Frameworks of task difficulty and complexity

Several frameworks of task complexity, especially for tasks that assess spoken language proficiency have been proposed; for example by, Candlin (1987), Robinson (1995, 2001), Skehan (1996, 1998), and Bygate (1996, 2001). Broadly speaking, these frameworks all attempted to categorize productive language tasks according to a common set of features and to understand how the manipulation of these features affects the language produced in response to tasks. It is this focus on task categorization that makes the frameworks of relevance to this research, even though they relate to speaking tasks, not writing.

Robinson's (1995, 2001) and Skehan's (1996, 1998) influential frameworks both took a cognitive approach to language learning (Tavakoli, 2009) and theorized that the difficulty of the task depends on a range of factors, such as linguistic complexity, cognitive complexity, and learner motivation.

Skehan (1998: 99) defined task difficulty in terms of the following three features:

1. Code Complexity (linguistic complexity, vocabulary load and redundancy and density);
2. Communicative Stress (time limits and time pressure, speed, number of participants);
3. Cognitive Complexity:
    a. Cognitive Familiarity (familiarity of topic, familiarity of discourse genre, familiarity of task)
    b. Cognitive Processing (information organization, information type, amount of computation, clarity and sufficiency of information given)

Skehan (1998), along with Skehan and Foster (2001, 2005) claimed that tasks that are more difficult will impact learners' spoken language accuracy, fluency, and complexity. That is, as task difficulty increases, the three features of spoken language production will be negatively affected and the accuracy or fluency or complexity of spoken language will be reduced. Skehan (1998) also claimed that as tasks become more cognitively demanding, learners will need to focus on the content of their responses and will be less able to focus their linguistic resources on the complexity and accuracy of their spoken language.

In contrast, Robinson (2001) viewed task complexity as distinct from task difficulty. In Robinson's framework, the difficulty of the task is dependent upon the learner's aptitude and motivation, which interact with task conditions that make demands on the learners. The task itself is analyzed in terms of its complexity, separate from the learners' aptitude. Robinson (2001: 29) defined task complexity as "the result of the attentional, memory, reasoning and other information processing demands imposed by the structure of the task on the language learner." Robinson (2005, 2007) claimed that more complex tasks will elicit spoken language that is more accurate and more complex than that elicited by less complex tasks. Conversely, more complex tasks will elicit less fluent spoken language.

The aspects of tasks that both Skehan and Foster believed influence the type of language elicited are the volume of information presented in the task and the number of tasks presented to the learner. In addition, Skehan's model indicated that the complexity of the syntax and lexis of the task may affect the language elicited. The familiarity of the topic presented in the task is also a key factor that affects the difficulty of the task and the type of language that it may elicit.

These frameworks provide a useful starting point to understand the distinguishing characteristics of constructed response tasks and hence, how they may be categorized. While both the frameworks focus on tasks that elicit spoken language, the frameworks do provide a detailed analysis of how tasks have an effect on the language that is elicited and, particularly in Robinson's model with the individual language learner. In addition, these frameworks provide a broad theoretical frame within which to situate an approach to prompt categorization that is directly relevant to the aims of this work: understanding the effect of a productive language task on the response elicited.

The following sections will report on studies that have focused explicitly on writing tasks and created their own approaches to writing prompt categorization, either as the primary or as a secondary aim of the study. This literature will be more directly relevant to the current study; to develop a systematic approach to categorizing the core characteristics of writing prompts than is possible with the more broadly drawn task complexity frameworks reported in this section.

3.2 Studies of writing tasks

This section narrows the focus of the review of literature, moving from the broad scope of task complexity frameworks to concentrate on previous studies that have explicitly attempted to classify writing tasks. This literature is directly relevant to one of the primary aims of this thesis, identifying the common characteristics of independent writing prompts. A small number of studies have had the sole aim of analyzing different writing tasks, mainly in academic contexts and this body of literature is valuable because of its explicit focus on task features. Most of these studies have not typically addressed

writing prompts used on ESL exams but, rather have examined writing tasks that are assigned to students in an English L1 college setting. However, such studies are certainly relevant to developing an understanding of the key characteristics of prompts on ESL exams because they illustrate the types of writing commonly asked for in the academic context.

Swales (1982) and Horowitz (1986a and 1986b) made valuable early contributions to the understanding of how writing tasks may be classified. Swales (1982) proposed categorizing examination prompts by the function of the instructional verb. For example, Swales stated that many examination questions require students to "describe", "compare and contrast", or "discuss" a range of topics. Focusing on the instructional verb provides some guidance to the examinee and the researcher with regard to the purpose of the examination question but this is far from a comprehensive approach to prompt categorization. Horowitz pointed out that questions with the same instructional verb may have quite different purposes and supported this claim with the following authentic examination questions that require a written response (Horowitz, 1986b: 107).

> Example 1: Describe the causes of the War of 1812.
>
> Example 2: Describe the technologies associated with horticulture and also those associated with agriculture.
>
> Example 3: Describe the relationship between population growth, urbanization, and the demographic transition.

Horowitz suggested that the writer is expected to respond to these questions in quite different ways; such as, an historical account, a list, and an abstract account of three theoretical principles. The language elicited by these tasks is likely to be quite different even though the tasks all contain the same instructional verb. Horowitz's conclusion was that Swales' categorization is insufficient to capture the complexities of writing tasks across different academic contexts.

Horowitz (1986a, 1986b), while influenced by Swales' approach, proposed a typology of essay prompts that consist of "a relatively small number of organization categories and a relatively large number of frames that are the linguistic realization of these categories," (Horowitz, 1986b: 109). This approach to writing prompt categorization was developed from a survey of prompts used at a single US university (Western Illinois University), where Horowitz collected a sample of 284 writing prompts from 29 courses in 15 different university departments. Horowitz's description (1986a) of the categorization process is brief and indicates that the categorization was created by the author reading the prompts and categorizing them according to the distinguishing characteristics of the sample; that is, an inductive approach to the categorization was taken and the resulting framework was not driven by any existing theory of prompt categorization or any broader theoretical position based in the second language writing or task difficulty literatures. Horowitz drew two sets of conclusions, one (1986b) that provided four broad categories of writing tasks, and a second (1986a) that described seven categories of finer grained task categories.

The broader set of four categories (Horowitz 1986b) for essay prompts was:

I. Display familiarity with a concept.
II. Display familiarity with the relation between/among concepts.
III. Display familiarity with a process.
IV. Display familiarity with argumentation.

Within these four categories, Horowitz gave multiple examples of the essay prompts that were analyzed in his study, which were based on the guiding principle of the function of the instructional verb.

Horowitz (1986a) described seven distinct categories of writing tasks:

- Summary of/reaction to a reading
- Annotated bibliography
- Report on a specified participatory experience
- Connection of theory and data
- Case study
- Synthesis of multiple sources
- Research project

Horowitz (1986a) went on to distinguish between the types of writing tasks found in the authentic academic environment and those sometimes presented in an ESL context. Horowitz claimed that:

> Generally speaking, the academic writer's task is not to create personal meaning, but to find, organize, and present data according to fairly explicit instructions. This is not to denigrate the concept of writing as personal discovery; it is merely to say that this type of writing appears to be absent outside of English composition, creative writing, literature, and ESL classes.

Horowitz's work is of interest because it attempted to categorize authentic writing tasks in the academic context, the same context that admissions tests such as the TOEFL, IELTS, and MELAB attempt to assess test takers' readiness to perform in. The resulting categories from Horowitz's work are broad but helpful as a classification of the types of tasks that writers in the academic context are likely to need to respond to. The shortcomings of the work are that the tasks were sourced only from a single university and the response rate to Horowitz's survey was only around 5% of university faculty, bringing into question the representativeness of the sample. Also, the resulting categories were derived only from the author's analysis of the tasks. The approach was inductive and guided by a single researcher's interpretation of a relatively small number of tasks. Horowitz provided an important step toward understanding academic writing tasks but the relatively simple approach taken, along with the restricted sample of tasks analyzed brings the extent to which the findings can be generalized into question.

ETS research into the writing test for the TOEFL 2000 project (Hale, Taylor, Bridgeman, Carson, Kroll, and Kantor, 1996) built on Horowitz's (1986a, 1986b) work. The aim of this work was similar to that of Horowitz in that it reported on a survey of academic writing tasks employed in the US and Canadian university context. In contrast to Horowitz though, the ETS study surveyed faculty at multiple universities at both the undergraduate and graduate level. Initially, 110 assignments were analyzed to develop a classification scheme for the writing tasks. The assignments were sourced from eight major research universities, both public and private in the US and Canada. The development of the initial classification scheme was inductive, similar to Horowitz's approach and based on the tasks analyzed, not

responses to the tasks. After the development of a tentative classification scheme, this was applied to all assignments by five of the researchers. The final classification scheme had six categories:

1. Locus of writing (in or out of class)
2. Length of product
3. Genre
4. Cognitive demands
5. Rhetorical task
6. Pattern of exposition

Hale et al (1996) supplied a great deal of supporting detail from the survey to exemplify the six categories. The locus category was intended to distinguish between assignments that are completed in class and those that are assigned out of class. Length of product was separated into five possible outcomes (Hale et al. 1996: 10)

- Non-extended discourse (around 75 words)
- Short extended discourse (1/2 to 1 page)
- Extended discourse (more than 1 page; up to 5 pages)
- Longer extended discourse (6 to 10 pages)
- Very extended discourse (more than 10 pages)

The genre category was used to explain what the written product looks like, such as an essay, the report of an experiment, or a case study. Ten separate genre categories were described in total. The category of cognitive demands was intended to "capture the level of thinking skills or intellectual functioning required to accomplish certain tasks, (Hale et al. 1996: 12). Categories such as cognitive demands are particularly problematic when relying on only an analysis of the tasks themselves. The cognitive demands the tasks make of writers is subjective and probably only a determination that can be made with the input of the students who respond to the tasks. Rhetorical tasks explain how the college instructor defines the forms or modes of discourse that the assignments take. The four categories identified were:

- Narration
- Description
- Exposition
- Argument

Patterns of exposition reflect the expected patterns that would be present in students' responses to the writing tasks, such as comparison/contrast, definition, and analysis. There were seven such patterns described in total.

Agreement rates between the researchers for the categorization decisions were reported and described as modest by the authors. Overall, the scale of the survey and number of assignments analyzed is impressive. However, the inductive approach, reliance on analysis of only tasks and not responses, and the modest agreement rates between the authors on the categorization decisions raise some questions about the merits of the categorization scheme. However, taken together, the work of Horowitz and the

ETS survey provide a good understanding of the types of writing tasks typically assigned in US and Canadian universities.

A similar study of IELTS writing prompts and academic writing tasks at two Australian universities was performed by Moore & Morton (2005). The aim of this work was comparable to that of Hale et al.; to compare the writing tasks employed on a high-stakes ESL test (knowledge telling tasks, per Scardamalia & Bereiter, 1987) with those writing tasks commonly found in authentic academic settings (knowledge transforming tasks, per Scardamalia & Bereiter, 1987). The authors analyzed 20 IELTS writing tasks and 155 university writing tasks at both the undergraduate and graduate level in several different disciplines. The four broad categories identified were:

- Genre
- Information source
- Rhetorical function
- Object of enquiry

Moore & Morton (2005) described 10 possible genre categories for writing tasks, the same number as in Hale et al., with these two sets of genre categories being very similar but not identical. Moore & Morton (2005: 51) go on to say that the IELTS writing tasks analyzed all represented just a single genre category (essay) compared to the 10 identified in the authentic college academic writing context.

The information source category referred to "the type of information that was to be used in the completion of a task (Moore & Morton, 2005: 51), most notably information from prior knowledge, primary sources, and secondary sources. While only 3% of the university writing tasks focused on the use of prior knowledge, all the IELTS tasks required only prior knowledge and no primary or secondary information sources. This finding highlights a common complaint about ESL writing tasks, even on tests of academic English, in that the ESL writing tasks fail to adequately represent the expectations of argumentation required in authentic academic contexts, a criticism echoed strongly by Perelman (2012).

Moore & Morton's (2005) categorization of rhetorical function was a little more detailed than that of Hale et al. (1996). Moore & Morton identified nine rhetorical function categories, as compared to the four of Hale et al. Moore & Morton provide additional categories, beyond the ones found in Hale et al., such as *evaluation*, *prediction*, and *recommendation*. The IELTS tasks relied heavily on the rhetorical function of evaluation and while this was also an important function found in authentic academic writing tasks, the range of rhetorical functions found in the authentic academic settings was considerably broader than those found on the IELTS writing tasks. The final category of object of enquiry operationalized the distinction between tasks that focused on specific, tangible subjects and those that were more abstractly or theoretically drawn.

Perhaps unsurprisingly, the findings of Moore & Morton's comparison of authentic university writing tasks and the tasks found on the IELTS indicated that the great diversity of writing tasks found in the academic context was not replicated on IELTS. As Moore & Morton (2005: 64) put it:

> It is our view that the form of writing being prescribed by the IELTS, on analysis, may have more in common with certain public forms of written discourse than with those of the academy. In particular, the emphasis placed on the spontaneous expression of opinion suggests such public

> nonacademic genres as the letter to the editor or newspaper editorial. Whilst practice in this type of writing will certainly contribute in a general way to students' literacy development (how to write coherently, grammatically, etc.), it would be a mistake in our view to see it as an appropriate model for writing in a university context.

While the aim of this thesis is to identify key characteristics of writing prompts on ESL tests and not in the authentic academic context, collectively the approaches of Horowitz (1986a, 1986b), Hale et al. (1996), and Moore & Morton (2005) provide a comprehensive review of the types of writing tasks students can expect to be faced with in higher education. Genre is the one category that is common to all three studies. Rhetorical task (or function) is another characteristic that is common to the work of Hale et al. and Moore & Morton and a distinguishing task characteristic within both approaches. The other categories described in the three studies are distinct from each other.

The main conclusion to be drawn from these three important studies is that genre and rhetorical task can distinguish between writing tasks in the higher education context. The other distinguishing characteristics differ based on the interpretations of the researchers and the specific tasks studied in the different contexts and institutions. These findings help illustrate that such classification schemes are quite subjective and depend on context and the perceptions of those coding the tasks. Without a more principled approach to task coding, beyond the inductive approaches undertaken in these studies, the resulting coding schemes will be based largely on the subjective interpretations of the researchers.

Kroll & Reid (1994) outlined a set of guiding principles for designing effective writing prompts in an ESL context. This work, in contrast to the three reported above focuses on ESL writing tasks and not the writing tasks found in the college context. Kroll & Reid (1994: 233) suggested that such writing prompts typically take one of three main formats.

1. A bare prompt
2. A framed prompt
3. A text-based or reading-based prompt

A bare prompt is described as one that "states the entire task for the candidate, (Kroll and Reid, 1994: 233). One example provided is:

> Many Americans like to participate in different sports. Write an essay in which you discuss the reasons why you do or do not like to play sports. Be specific.

A framed prompt provides a situation or set of circumstances along with a task that is based on the interpretation of the frame. An example cited by Kroll & Reid is:

> Some people feel that using animals for food is cruel and unnecessary, while others feel that it is necessary for people to eat meat and that the production of animals for food can be done without cruelty. What is your position on the issue of whether people should use animals for food? Discuss the strengths and weaknesses of both positions and use concrete examples when you explain your point of view.

The distinction drawn between bare and framed prompts is an important one for independent writing prompts. It helps distinguish between different independent writing prompts in terms of both length

and complexity and provides some clear guidance for an approach to independent prompt categorization.

The final prompt type is equivalent to a reading-to-write prompt, or integrated writing prompt. The test taker is required to read an authentic reading passage and then write a response based on the passage. Within these three prompt types, Kroll & Reid listed six variables that prompt designers need to consider when crafting a writing prompt.

1. The writing situation (contextual variables)
2. The subject matter (content variables)
3. The wording of both the prompt and the instructions (linguistic variables)
4. The task(s) (task variables)
5. The rhetorical specifications (rhetorical variables) and
6. The scoring criteria (evaluation variables)

Points two to six are of particular relevance to this thesis as these points specifically focus on the characteristics of the prompt itself. Within content variables, Kroll & Reid (1994) emphasized the importance of constructing writing prompts that are based on topics that tap into the background knowledge of the test taker. Prompts should be based on topics that are familiar to as many test takers within the test population as possible so as to not unfairly advantage those who are more familiar with a certain topic than others. The content of the prompt should also be equally accessible to all test takers regardless of cultural background. Prompts that prove to be inaccessible or confusing to test takers from certain cultures should not be used operationally.

Under linguistic variables, Kroll & Reid (1994) called for unambiguous wording in test instructions. Beyond the instructions, Kroll & Reid (p.237) stated "the prompt itself should be worded in a way that is transparent and easy to interpret in terms of vocabulary and syntax, avoiding any possibility of ambiguity, be it linguistic or cultural."

Kroll & Reid (1994) also called for prompt designers to consider the number of tasks that are set for the test taker within the prompt. This number needs to be determined in relation to the time allowed for the writing test. The need for an accessible topic and a restricted number of tasks are emphasized for a timed writing test.

Kroll & Reid's guidance regarding rhetorical variables lacked the clarity of their work on the other variables. The main aspects of rhetoric they considered are the expected voice of the response, the expected audience, and the expected rhetorical structure of the response based on the instructional verb in the prompt. Kroll & Reid cautioned against making the rhetorical demands excessive in the prompt and the instructions as they believe that such prompts may elicit "repetitive essays that can lead to difficulties in scoring," (Kroll & Reid, 1994:239).

Overall, the guidance provided by Kroll & Reid (1994) on writing prompt design allows for an understanding of characteristics of writing prompts that are common across a variety of prompt types and provides useful guidance for those who wish to reduce writing prompts to their core common characteristics.

Finally, Hamp-Lyons & Mathias (1994) undertook an investigation of writing prompt characteristics that is very relevant to this research as it focuses on the same testing context; the MELAB exam from The University of Michigan. They investigated the difficulty of 64 MELAB writing prompts, using expert judges to determine the relevant difficulty of the prompts. Hamp-Lyons & Mathias identified two main distinctions between the 64 prompts that they examined. These were the expected response mode that the prompts elicited (either expository or argumentative responses) and the orientation of the prompts (whether the prompts were publically oriented or privately oriented). The authors identified the prompt length, syntactic complexity of the prompt, and the expected audience as being relatively stable for the MELAB and not being relevant for the prompt categorization framework. The two primary distinctions (response mode and orientation) were used to create five task-type categories, (Hamp-Lyons & Mathias, 1994: 55).

1. Expository/private
2. Expository/public
3. Argumentative/private
4. Argumentative/public
5. Combination of two or more of the other four types

The expert judges (MELAB raters and ESL writing scholars) hypothesized that the expository/private prompts would be the easiest for test takers and that argumentative/public prompts would be the most difficult. However, an analysis of the scores awarded to written responses (n=8,583) to the 64 prompts showed that the opposite to the hypothesized hierarchy of prompt difficulty was the case. The average score awarded to the responses to the expository/private prompts was the lowest and the highest average score was awarded was to the responses to the argumentative/public prompts. Hamp-Lyons & Mathias speculated that the raters may be compensating for anticipated prompt difficulty when they evaluate the responses to the different prompt types. The authors concluded by claiming that expert judges are not accurate predictors of prompts that will be difficult or easy for test takers, a finding that is strongly supported in their data. While the predicted difficulty of the different prompt types identified by Hamp-Lyons & Mathias may be inaccurate, the categorization system they developed is of value when finalizing a categorization framework for independent writing prompts.

The literature described above shows that there are many different ways to categorize writing prompts. It is challenging to see a common thread to the findings from the investigations of writing tasks described here and there is no conclusion easily drawn regarding how to best identify the core common characteristics of writing tasks. This is partly because these works have different foci but it is also because the tasks analyzed in each of the studies were quite different. The tasks analyzed in the work of Horowitz, Hale et al., and Moore & Morton were those found in authentic college content based courses. In contrast, the prompts described by Kroll & Reid (1994) and Hamp-Lyons & Mathias (1994) were ones found on ESL writing tests.

The work of Horowitz (1986 a, 1986 b), Hale et al. (1996), and Moore & Morton (2005) helps readers to understand the range of writing tasks used in the academic context that ESL writing tests must attempt to assess readiness for. While the categorizations reported in Horowitz (1986a, 1986b), Hale et al (1996), and Moore & Morton (2005) are of great relevance to the determination of the validity of ESL writing

tests, the variables identified by Kroll & Reid (1994) and the categorization employed by Hamp-Lyons & Mathias (1994) are more applicable in determining the core common characteristics of independent writing prompts. These are the bare and framed prompts in Kroll & Reid's work. Hamp-Lyons & Mathias (1994) provided a helpful distinction in prompt characteristics with the categories of response mode and orientation; however, identifying just two categories may prove to be overly simplistic and fail to allow for a full understanding of the core characteristics of writing prompts. The work of Kroll and Reid (1994) and the identified prompt variables of *content*, *linguistic*, *task*, and *rhetorical* variables is promising in providing a little more nuance to the categorization of writing prompts. Interestingly, one task feature that is common among all the work reported in this section (with the exception of Horowitz) is that of rhetorical task, or response mode in the terminology of Hamp-Lyons & Mathias (1994). Whether a task elicits an argumentative, expository, or narrative response is the one characteristic that has been identified as distinguishing between writing tasks by almost all studies. This is one prompt characteristic that will be vital to consider when creating the prompt categorization framework for this thesis.

One serious issue with all these studies is the lack of attention paid to the responses to the tasks and how the intended audience (the students or test takers) interpreted the tasks. The writing tasks have been categorized based on the researchers' (not the test takers') interpretations of the tasks (Hamp-Lyons & Mathias also collaborated with ESL raters). While these are valid and important interpretations, the findings would benefit from having been triangulated with an analysis of authentic responses to the tasks and by gaining an insight into how the intended audience (the student or test taker) interprets the writing tasks. Without such evidence, the categorizations that help us understand the key common characteristics of writing tasks will continue to be based on the interpretations and assumptions of expert judges that the work of Hamp-Lyons and Mathias showed so clearly to be, at times misguided. In contrast, this thesis will not only consider the approaches adopted in the relevant literature, but will also make use of evidence provided by test takers and by essay raters who read several responses to each of a number of different prompts. It is hoped that this triangulated approach to prompt categorization will offer some fresh insight to the second language writing field with regard to the characteristics of writing prompts that best distinguish between different prompts.

3.3 Writing prompt categorizations in studies of prompt effect

The studies reported above were directly focused on descriptive accounts of productive language tasks; either frameworks of speaking tasks (Section 3.1) or analyses of how writing tasks have been categorized (Section 3.2). Many more studies have included an approach to writing prompt categorization, typically in work that investigates how manipulating certain aspects of writing prompts affects test taker performance on writing tests. This section will review how writing prompts have been categorized in studies where the direct focus is not on the prompt categorization but on the written language elicited by the prompts. These studies have already been described in detail within Chapter 2 (Sections 2.1 and 2.2) so the approaches to prompt categorization in these studies will be summarized in the table below.

Table 3.1: Summary of prompt categorization approaches within prompt effect studies

Study	Prompt categorization	Population	ESL writers
Greenberg (1981)	Experiential/cognitive demand	College students	No
Quellmalz et al. (1982)	Discourse mode/response mode	High-school students	No
Brossell & Ash (1984)	Personal/neutral & imperative/neutral	College students	No
Smith et al. (1985)	Complexity & length of topic structures	College students	No
Hirokawa & Swales (1986)	Simple/academic	College students	Yes
Spaan (1990)	Rhetorical specification/subject matter	MELAB takers	Yes
Peyton et al. (1990)	4 different task genres	6th grade students	Yes
Brown et al. (1991)	Integrated/independent tasks	College students	Yes
Cumming et al. (2005)	Integrated/independent tasks	TOEFL takers	Yes
O' Loughlin & Wigglesworth (2007)	Quantity and presentation of task information	College students	Yes
Kuiken & Vedder (2008)	Quantity of task information	College students	No
Ong & Zhang (2010)	Time/amount of detail of input	College students	Yes
Lim (2010)	Topic domain/task constraint/rhetorical task/prompt length/grammatical person of response	MELAB takers	Yes

As may be apparent from the numerous studies of prompt effect reported above, there have been a wide variety of approaches adopted to prompt categorization. Some of these studies are more relevant to this thesis than others. The studies that contrast independent writing prompts with integrated prompts are not especially relevant to a categorization of only independent prompts, which is the focus of this thesis. The aim of this study is to establish the effect of the characteristics of independent writing prompts (bare or framed in Kroll & Reid's terminology) on the written product and test taking processes of second language writers. The distinctions between independent and integrated writing prompts are considerably broader and more substantive than those between different independent writing prompts. Therefore, the studies that focus on the distinction between independent and integrated writing prompts (Brown et al., 1991; Cumming et al., 2005) do not provide sufficient relevant guidance toward a meaningful set of independent prompt characteristics. Two other studies (Peyton et al., 1990; O' Loughlin & Wigglesworth, 2007) also lack relevance to an analysis of independent writing prompts. Peyton et al. (1990) focused on prompts that elicited different genres of writing whereas independent prompts typically elicit only letters or essays so there is little merit in using genre as an identifying

category. Likewise, O' Loughlin & Wigglesworth (2007) utilized a task type that is not typical of an independent writing prompt, that is a prompt that asks writers to interpret a graph or chart and to compose a response based on information presented graphically. This is not to question the validity of such tasks but these tasks are not the focus of this thesis.

The review of literature of prompt categorization shows that several different approaches to categorization have been undertaken. Some trends that emerge are:

1. The complexity of the input, either in terms of quantity/length of the input or the linguistic difficulty of the input has been used to categorize prompts in several of the studies with either a binary distinction or a continuum of complexity being used to distinguish between prompts.
2. The distinction between personal and non-personal responses has been used to make a binary distinction between different prompts in several studies.
3. Different modes of response (narrative, expository, argumentative) or genres of writing (journal writing, letter, essay) have been used to distinguish between different types of prompts.
4. The distinction between independent and integrated writing tasks has been the subject of some attention more recently in the literature

The first three of these approaches provide a valuable summary of the main writing prompt characteristics that have been used to categorize writing prompts in the second language writing assessment literature. These prompt characteristics are all considered carefully in the initial phase of prompt categorization in this study, described in Chapter 4 (see Section 4.2.2). Some of the studies described in this chapter are of particular relevance to the aims of this work. Hamp-Lyons & Mathias (1994), Spaan (1990), and Lim (2010) all worked with writing prompts from the MELAB test program and the findings of these studies must be considered carefully when developing a categorization framework for the assessment context that is the focus of this research.

The most detailed approach to prompt categorization was that developed by Lim (2010) who identified five distinct prompt characteristics that could distinguish between different independent writing prompts. However, Lim's categorization suffers from the same shortcomings that apply to many works described in this chapter. It is based only on the observations of the researcher and derived from the researcher's interpretation of the prompts themselves. It does not take into account any analysis of the responses to the prompts and does not call on the people who engage with the prompts; the test takers.

While Lim's framework is the most comprehensive one seen (for writing prompts used on ESL exams) in the second language writing field, it cannot be held up as definitive until the findings are triangulated from sources beyond a single researcher. It is these concerns with how previous categorizations have been arrived at that drives an alternative approach taken in this work; an approach that does not rely on expert judgments but also draws on the views of experienced raters, a review of prompt responses by these raters, and the input of authentic test takers as to which aspects of the prompts are important and distinguishing.

3.4 Rationale for focus on independent writing prompts

After the review of relevant literature, this presents an opportunity to clarify why the focus of this thesis is on independent writing prompts. The work of Hale et al. (1996) and Moore & Morton (2005) was important in moving the writing assessment field toward a research focus on integrated writing tasks. Hale et al. and Moore & Morton had demonstrated that writing tasks on high-stakes EAP exams (the TOEFL and IELTS) only partially represented the construct of writing as demonstrated in the summary of writing tasks that college students were routinely expected to engage with. The development of integrated writing tasks such as the reading-to-write or listening-to-write tasks now operational on the iBT TOEFL was a step taken to make these tasks more construct representative of academic writing.

The work that followed in the field of writing assessment (Cumming et al., 2006; Plakans, 2009) focused on how different integrated tasks influenced the written language elicited, with an emphasis on the effect of the spoken or written input that must be processed by the test taker. Though similar task types have been in use since the 1930s on CPE, (Weir, Vidaković, & Galaczi, 2013) this was an understandable direction for the field to take, with integrated tasks now being employed on a high-stakes EAP assessment (iBT TOEFL). However, the new interest in integrated writing tasks came about before there was any clear consensus on the effect of different independent prompt characteristics. At the core of independent writing prompts are the instructions that direct the test taker how to respond and pose specific questions to address. These features are also present in integrated tasks. Indeed, any constructed response task must require test takers to take a certain course of action in their response. It is these core aspects of writing prompts that are of most interest to this research. The effect of the lengthy input material in an integrated task is important for the field to address. But without an understanding of how the task requirements in the prompt affect the response, we will only have a partial understanding of the effect of the prompt characteristics on the response elicited.

3.5 Summary

It is challenging to generalize the findings of the literature described in this chapter. There is no consensus in the literature regarding the most appropriate categorization framework for independent writing prompts. The task complexity frameworks were intended for use with speaking tasks and not those found on second language writing assessments. The dedicated work on writing task categorization was most commonly performed in North American and Australian universities and the resulting task categories are relevant to the writing tasks found in undergraduate and graduate degree programs but are less applicable to the ones used on the writing sections of ESL exams. The prompt effect studies described are also situated in several different learning contexts, some that feature second language writers and several that do not. They have some serious methodological shortcomings, most commonly a reliance on an inductive approach from only the researcher(s), as stated above (p.49). There is a broad body of work in this literature but there are as many differences in the contexts and findings as there are similarities.

All of the categorization frameworks described were developed from an inductive analysis of the writing prompts themselves. With the exception of Hamp-Lyons & Mathias (1994) who sought the input of

raters, virtually none of the studies reported above described any analysis of responses to the writing prompts or any contribution from raters and test takers in developing the prompt categorization. This reliance on the insights of the researcher or research team to prompt characteristics presents a risk that the resulting categorization may be narrowly drawn. While taking into account the findings of the literature reported in this chapter, the approach to prompt categorization used in this work will also be based on a reading of responses to a range of writing prompts and the input of raters and test takers. It is hoped that this approach will assist in creating a prompt categorization framework that builds on the existing literature but also makes an original contribution by taking into account a broader range of views than those of previous studies.

It is important to keep in mind why a detailed prompt categorization framework (one that realistically captures the key characteristics that distinguish between different prompts) is necessary in order to investigate prompt effect. To determine the effect of different prompts on the writing sample produced, it is important to understand the generalizable features of independent prompts by analyzing their distinguishing characteristics. If the writing prompts can be categorized into distinct types, it will be more straightforward to understand which of these characteristics may be responsible for any prompt effect detected. That is, without such a categorization, any prompt effect detected would be challenging to interpret in terms of the prompt characteristics that may have contributed to the effect.

An aim of this work is to make a contribution toward a better understanding of how to establish task equivalence across test forms. In order to do this, it will be necessary to understand which features of the prompt must be controlled for equivalence to be established. Without that understanding, task equivalence will remain elusive. If the prompt categorization is too simplistic then the features of the prompt that may be responsible for prompt effect may be impossible to identify. If the prompt categorization is too elaborate, it may be impossible to operationalize the framework in a practical study of prompt effect and it may also be challenging to detect significant effects of specific prompt characteristics. Hence, the prompt categorization framework needs to identify the characteristics of the prompts that meaningfully and repeatedly distinguish between different types of independent writing prompts.

Institutions, such as exam boards and intensive English programs that administer writing assessments, and individuals who create writing prompts for high-stakes exams (or anyone who has an interest in the fairness of test scores or how to interpret test scores) would benefit from an understanding of how to make writing prompts equivalent. But this equivalence can only be established if there is an awareness of the characteristics of the prompt that make one prompt equivalent to or different from another.

Chapter 4 – Main study: materials and methods

Chapters 2 and 3 identified three major gaps in the research literature on the effect of independent writing prompts on high-stakes assessments:

- Lack of agreement regarding the key common characteristics of independent writing prompts
- Limited understanding of the relationships between prompt characteristics and the textual features of responses
- Lack of evidence about the effect of the prompt on writing process, reflecting an over-reliance on quantitative analyses of written products

4.1 Research questions

Specifically of interest in this work, is how test takers respond when presented with different independent writing prompts, how they utilize the writing prompt during the test (both prior to and during the act of composing), and how different independent writing prompt characteristics affect the nature of the written produced by the test taker. Hence, the research questions addressed in this work are as follows.

1. What are the characteristics that distinguish between independent writing prompts?

2. How do these characteristics affect the test-takers' final written product?

3. How do these characteristics affect the test takers' test taking processes?

This chapter will describe the mixed methods that were utilized to answer the first two of these research questions. The third research question will be addressed in Chapter 7. First, the steps taken to categorize the core characteristics of independent writing prompts will be reported. Second, the methods employed to investigate the relationships between different prompt characteristics and textual features of the written responses will be described. This will include an account of how the written responses were analyzed, according to a range of key indicators of second language writing proficiency. The aim of this chapter is to present a thorough account of how the research was conducted so it may be fully replicable by others.

4.2 Categorizing writing prompts

In Chapter 3, I reviewed previous attempts in the literature to identify key writing prompt characteristics and to formulate prompt categorization frameworks. Only a small number of prompt characteristics were common to multiple studies. The most commonly identified characteristics included the complexity of the input, the distinction between personal and non-personal responses, the different modes of response, and the genre of the response. Methodological shortcomings described in Chapter 3 (summarized in Section 3.4) mean that existing frameworks cannot be used, unmodified, for this study.

However, they do offer a helpful starting point for the creation of a new prompt categorization framework.

Previous attempts to identify the common characteristics of writing prompts were driven by an inductive approach and relied on a prompt review by the researcher(s). There was little evidence of any analysis of responses to the prompts or input from experienced raters or test takers, both of whom may be able to provide insights into features of the prompts that help distinguish one from another. In this section, a different approach to identifying a set of prompt characteristics will be described, one that begins with the findings from the literature but one that then takes into account a reading of authentic responses to independent writing prompts, and also considers the views of experienced raters and test takers. It is hoped that this approach will allow for a triangulation of different perspectives on the characteristics of independent writing prompts.

4.2.1 Data collection: Sourcing writing prompts for categorization

This section reports on the sourcing and analysis of a number of writing prompts from an ESL testing program. The writing prompts were sourced from the *Michigan English Language Assessment Battery* (MELAB), a test of English for Academic Purposes. I was granted permission to collect data on this test program while a full-time employee of The University of Michigan. I completed compulsory University of Michigan research ethics and compliance courses within the university's Program for Education and Evaluation in Responsible Research and Scholarship (PEERS).

The MELAB is a test of advanced level English proficiency and is typically used to determine readiness for academic admissions or professional licensure. According to the MELAB Technical Review (CaMLA, 2015):

> The MELAB is intended for adult nonnative speakers of English who are seeking admission to colleges and universities where the language of instruction is English, or for professional purposes. Consequently, the content and tasks are drawn from the formal and informal communication contexts a college or university student might encounter, as well as general occupational or office settings. These include conversations between friends and service encounters as well as the interactions and inputs that might be expected in seminars and lectures. The MELAB is a multilevel exam, covering a range of proficiency levels on the CEFR (Council of Europe, 2001) from B1 to C1; test takers at the B1 and B2 levels are considered independent users of English, and test takers at the C1 level are considered proficient users of English.

Lim (2009: 22) argues that "the MELAB takes a trait perspective to construct definition, and defines its writing construct to be the ability to produce a composition of some length that is appropriately organized and developed and that evidences control over different aspects of the English language. This narrowed down writing construct excludes a number of possible types and genres of writing, but is also sufficiently broad as to cover a type of writing often seen and employed in educational and professional contexts.

The MELAB is a good source of independent writing prompts because of the broad variety of topics covered and the range of perspectives that test takers are encouraged to take. The MELAB is also a

standardized, high-stakes assessment that is designed and developed according to rigorous specifications and guidelines. Additionally, each MELAB test form (a unique test form is administered every month) provides test takers with a choice of two prompts, each intended to provide an equivalent challenge for the candidates and to elicit comparable written language skills. The writing section is the first part of the MELAB and is followed by compulsory listening, grammar, cloze, vocabulary, and reading sections (see Table 4.1 for details).

Table 4.1: Structure of the MELAB

Section	**Time**	**Description**	**Number of items**
Writing	30 minutes	Write one essay based on two prompt choices	1 task
Listening	40 minutes	Part 1 – Listening Questions	18
		Part 2 – Listening Dialogs	22
		Part 3 – Listening interviews	20
Grammar	80 minutes	Select word or phrase to restore correct grammatical meaning to sentence	32
Cloze		Two passages with multiple words or phrases deleted – test takers restore intended meaning to passages	24
Vocabulary		Select word or phrase to restore correct meaning to a sentence	31
Reading		Four reading passages, each followed by several reading comprehension questions	23

In the Writing section of the MELAB, the test taker is required to produce a single essay in response to a short prompt and test takers must select from two prompt choices. The essay is scored by two trained raters using the MELAB Rating Scale (see Appendix 2). According to Jung, Crossley, & McNamara (2015: 5) "the descriptors of the rating scale address various aspects of writing such as topic development, text length, breadth and appropriateness of lexical choice, adequate morphological and syntactic control, cohesion, and accuracy of spelling and punctuation."

Writing prompts are paired on MELAB test forms by ensuring that the two prompts offer a genuine choice of topics. Prompts that are situated in the same domain (educational, occupational, public, personal) or present similar topics are not paired together.

There is also an optional speaking test that may be taken if required by the academic institution or licensing body that candidates are taking the MELAB for.

4.2.2 Data collection: Identifying prompts for categorization

A total of forty different writing prompts were initially identified from the MELAB database for analysis. They were drawn from numerous different administrations of the MELAB over several years and covered a broad range of topics. These prompts were selected based on the findings of several studies on prompt categorization (see Chapter 3) because they displayed variety with respect to characteristics identified in Chapter 3:

- The complexity of the input, either in terms of quantity/length of the input or the linguistic difficulty of the input
- The distinction between personal and non-personal responses
- Different modes of response (narrative, expository, argumentative)

Some of the characteristics described in Chapter 3 could not be included. Genre of response could not be addressed using MELAB prompts as only essays are elicited. Some studies reported in Chapter 3 (Hale et al., 1996; Horowitz, 1986b; Moore & Morton, 2005) identified genre of response as a distinguishing characteristic of writing tasks but these were chiefly the tasks employed in higher education contexts and not those found on ESL writing tests.

The following sections elaborate on the details of how the prompt categorization approach was finalized. Briefly, the following steps were undertaken:

i. The initial stage of prompt categorization was based on the findings of the literature review described in Chapter 3. A number of MELAB writing prompts were identified that could be differentiated by the prompt characteristics shown to be important in the literature.
ii. Next, responses to the prompts of interest were read by the researcher to explore whether the responses were helpful in further identifying differences in the prompts.
iii. After that, these responses were also read by five experienced MELAB raters to verify whether the researcher's insights into the responses were shared by raters.
iv. Finally, to capture the perspective of the people who respond to the prompts, seven test takers were asked about the characteristics of writing prompts that made a meaningful difference to the difficulty or complexity of the writing prompts.

The previously identified categories of prompt characteristics (Chapter 3) informed the first stage of classification and selection of the prompts. Some of the initial 40 prompts were discarded as a result of overlap in key characteristics with other prompts; for example, many prompts elicited argumentative responses, asking for the test taker to provide a supported opinion. While it was important to retain some of these prompts as they are very representative of many MELAB writing prompts, and response mode is prevalent in the literature as a distinguishing prompt characteristic, some of these prompts could be discarded due to prompt characteristic overlap.

The aim of this initial screening of the prompts was to narrow the sample by removing prompts that did not differ from others based on the key prompt characteristics identified in the literature. This initial round of analysis, driven by the findings from the literature reported in Chapter 3 left 10 MELAB writing prompts that differed by complexity of the input (linguistic difficulty and length), the expected person of the response, and by the response mode. During this initial screening, it became apparent that another

prompt characteristic that could distinguish between the prompts was the number of tasks. This is a prompt characteristic that has been identified in the literature (Kroll & Reid, 1994) as a distinguishing characteristic between prompts. Most of the prompts within the MELAB bank set the writer only one or two tasks to respond to (commonly, give an opinion[1] and support it with reasons[2]), but some prompts within the bank contained multiple tasks (as many as six) and this characteristic was identified as potentially being worthy of consideration within the prompt categorization framework.

4.2.2.1 Data collection: Analyzing written responses to the prompts

In order to refine this initial classification, 10 written responses to each of the 10 prompts that had been retained (100 responses in total) were examined by the researcher. The objective was to explore whether there were any markedly different patterns in the responses to different prompts and to generate hypotheses about which prompt characteristics might be associated with these substantive differences in written products. This is not an approach that has typically been undertaken in the literature reported in Chapter 3. Previous prompt categories have been derived from an analysis of the prompts, not from the responses.

The addition of a phase reviewing the responses to the prompts was motivated partly by the findings of Hamp-Lyons & Mathias (1994), who showed that supposedly expert judgments of prompt characteristics may sometimes be less than expert, and also partly by the researcher's experience of pretesting writing prompts. In the MELAB program, writing prompts are pretested on a representative sample population in ESL programs before any prompt is used operationally. Responses to the trial writing prompts are reviewed qualitatively and the decision to approve, revise, or reject the trial prompt is made based on a reading of the responses along with a review of feedback from the participating students and teachers. Reviewing the pretested responses to prompts proved to be an effective way of identifying prompts that may not elicit appropriate responses on an operational test. Hence, a careful reading of responses to prompts was seen as a potentially useful addition to the approaches to prompt categorization reported in the literature reviewed in Chapter 3.

In order to identify differing trends in the sample of responses to ten prompts, an approach termed 'rich feature analysis' by Barton (2004) was adapted for this study. Barton claimed this approach is useful "in the analysis of academic discourse, particularly in the analysis of student writing and in the comparison of texts written by inexperienced writers (students, or new members of a disciplinary community) and experienced writers (established members of a disciplinary community)," (p.67). Although the purpose of this study is rather different, Barton's approach is of relevance because of its focus on identifying patterns within written language.

The relevant procedures outlined by Barton are as follows:

1. Select an initial corpus.
2. Identify salient patterns
3. Determine "interestingness"
4. Select a study corpus.

These steps were adapted to investigate the sample of 100 MELAB essays and determine salient patterns within them. The responses to the ten prompts in the initial corpus were read for their "interestingness" or whether there were salient patterns within the responses that differentiated them from responses to other prompts. For example, responses to some prompts were seen to be richly descriptive with a great deal of personal content, while responses to other prompts were typically made up of formulaic five-paragraph essays. These were the types of patterned distinctions in responses that were read for using Barton's approach.

The prompts that elicited the responses that stood out from the others in the mini-corpus could be distinguished based on the following prompt characteristics.

- Total length
- Number of sentences
- Number of words per sentence
- Number of tasks to respond to (rhetorical cues)
- Expected response mode (narrative, argumentative, or expository)
- Expected person of response

These prompt characteristics were consistent with the ones highlighted in the literature reviewed in Chapter 3. The first three points reflect prompt complexity. The final three points reflect the number of tasks, response mode, and expected person of the response (personal or non-personal). Up to this point, following an analysis of the prompts based on findings from the literature and then a reading of responses to the prompts, the prompt characteristics that distinguished between both prompts and responses were in line with findings from previous prompt categorization frameworks (see p.49).

The five prompts identified as eliciting responses with the greatest variation were as follows.

1) In some countries, such as the United States, the standard work week is 5 days, 8 hours a day. Some companies are considering changing their work week to 4 days, 10 hours a day. What are the advantages and disadvantages of the two options? Which do you think would be better for society? Give reasons to support your position.

2) In order to prevent illnesses such as food poisoning, the government enforces rules about how restaurants can store and serve food. Some local governments are now considering banning foods that might have long-term negative effects, such as foods with high amounts of fat. Do you think it is a good idea for governments to tell restaurants what they can and cannot sell based on possible long-term health consequences? Give reasons to support your position.

3) In everyone's life, some days are particularly memorable. Describe what happened on one such day that you have experienced in your own life. What made it memorable and why? Include specific details such as where you were, who else was there, and so on, that will convince the reader how memorable this day was.

4) Describe a situation in your life in which you wanted something very much, but regretted it (were sorry) after you got it.

5) Some people think neighbors should be friendly and sociable. Others do not want to be bothered by their neighbors. Explain what you think a "good neighbor" is.

A summary of the prompt characteristics initially identified is presented in Table 4.2 below.

Table 4.2: Initial prompt categorization

Prompt	**Length**	**Number of sentences**	**Words per sentence**	**Rhetorical cues**	**Response mode**	**Expected person of response**
1	59 words	5	11.8	6	Argumentative	Non-personal
2	74 words	4	18.5	2	Argumentative	Non-personal
3	54 words	4	13.5	6	Narrative	Personal
4	22 words	1	22	1	Narrative	Personal
5	27 words	3	9	1	Expository	Non-personal

The six categories shown in Table 4.2 are consistent with those present in the literature reviewed in Chapter 3. The term rhetorical cues refers to the number of sub-tasks a writer must respond to in order to fully address the prompt. The number of tasks has also featured in previous approaches to prompt categorization, although the specific counting of cues within a single prompt is most relevant to the testing context used in this study. Counting tasks, or in this case cues is consistent with previous approaches to prompt categorization (Kroll & Reid, 1994). The final two categories of response mode and expected person of response have also been utilized in several approaches to writing prompt categorization (Brossell & Ash, 1984; Hamp-Lyons & Mathias, 1994; Lim, 2010; Quellmalz et al., 1982), as reported in the review of literature in Chapter 3.

4.2.2.2 Verifying the prompt categorization

In order to verify whether other readers could identify similar trends in responses to these prompts (Table 4.2), five experienced MELAB raters agreed to read responses to each of the prompts presented above. They were asked to read the essays carefully and to consider whether there were any clear trends among the responses to particular prompts. The task was left deliberately broad so as not to

influence the readers to any particular conclusion but rather for the readers to be able to independently read for differences among the responses to the five prompts.

After the reading was completed by the raters, the researcher met with all five raters to discuss their views. These discussions were not grounded in a formalized approach to interview data collection, but rather were based on the shared experiences of a small team of experienced raters. All five raters had been rating MELAB responses for at least two years and two of the raters were also experienced MELAB rater trainers. These shared collective experiences of working with MELAB essays and the MELAB rating scale provided a rich base of experience to aid in identifying salient patterns (Barton, 2004) in the responses. The raters agreed that the responses to prompts three (memorable days) and four (regrets) (see p.57) were substantively different to those of the other prompts. The textual features of the responses to prompts three (memorable days) and four (regrets) that were consistently identified by raters were:

- Use of low-frequency vocabulary relevant to prompt
- Flow of response does not depend on use of mechanical transition markers
- Use of concrete supporting examples

After this first round of discussion, two of the initial five raters were given another set of 25 compositions to read (a further five essays per prompt). The reasons for continuing with only two raters for the second stage of reading were twofold; firstly, it would have been challenging to discuss 50 compositions in detail with all five raters. Working with a smaller group helps the discussion to retain focus and clarity, without individuals becoming confused or details of previously discussed essays being forgotten. Second, the two raters who continued for the second reading were the ones who had made the most insightful comments at the first meeting. The purpose of the second round of reading was to look in greater detail at the responses to the prompts that were identified as eliciting notably different responses in the first round. In addition, the remaining raters were asked to look again at responses to the other prompts to read for any distinguishing features that may have been missed during the first round of reading.

During the discussion of the further set of essays, the raters confirmed that prompts three and four elicited responses that were lexically richer (used a wider range of vocabulary) than responses to other prompts and also built cohesion with less of a reliance on mechanical transition markers. The features of the responses to prompts three and four agreed upon by the raters were as follows.

Language Related

- Use of low-frequency vocabulary relevant to prompt
- Use of appropriate collocations
- Richly descriptive adjectives and adverbs
- Flexible use of language to avoid repetition
- Use of appropriate idiomatic language

Content / Discourse Related

- Personal content that draws reader in
- Development flows
- Flow does not depend on use of mechanical transition markers
- Use of specific supporting examples
- Arguments developed in depth
- Avoids repetition
- Creates connection between reader and writer
- Creates imagery
- Able to build and sustain momentum

Some compositions written in response to prompts one (standard work week) and five (neighbors) were identified by the raters as containing linguistic and content/discourse features that were quite different from the features in the responses to prompts three (memorable days) and four (regrets). The distinguishing features of the responses to prompts one and five included repetitive content, a lack of clear structure and organization, a lack of personal or descriptive content, and use of generic vocabulary that lacked specificity. The raters also noted a tendency toward a reliance on a five-paragraph structure in some responses to these two prompts.

4.2.2.3 Finalizing the prompt categorization

The discussions with MELAB raters suggested that narrative responses, situated in the personal domain were notably different from responses to other prompts. Argumentative responses that were not in the personal domain exhibited distinctive features of cohesion (there tended to be more of a reliance on mechanical transitions markers, such as *first, second, next*). They also tended to be structured in a more rigid way (reliance on standard five-paragraph essay response) than the responses to other prompts. Hence, two characteristics of prompts that appeared to be important for categorization, (because they may elicit linguistically and discoursally distinctive responses) were response mode and expected person of the response.

To finalize the approach to prompt categorization, MELAB test takers were interviewed to gain an insight into the views of those who interact with the prompts. Individual test takers were interviewed after they had completed a live administration of the MELAB and were asked for their reasons for selecting the prompt they responded to and for their reasons for not selecting the other prompt available to them. Interviews were conducted following two different administrations of the MELAB. Seven test takers (four male and three female) were interviewed at the Toronto MELAB center. They ranged in age from 19 to 37 and had four different first languages (Tagalog, Arabic, Malayalam, and Punjabi).

These test taker interviews revealed that the length of the prompt (the number of words or the number of sentences) was not a significant factor for the interviewees when they selected a prompt. None of the test takers interviewed reported longer prompts being any more difficult to understand or respond to than shorter prompts. They were seen as equivalent by the test takers. This is not consistent with the

literature review (Chapter 2 – see Section 2.3), which indicated that the complexity of the input, in terms of length and linguistic complexity may affect the written product.

> Interviewer: OK. Just one more question about Topic A. Was there anything that was confusing or anything that was not clear about that?
>
> 10232010-03 (L1 Tagalog): Oh, everything is clear. I love it, you understand. Because to me it's immoral this kind of thing and you can do it free but it must be at a particular time, not just anywhere. So, that's it.
>
> Interviewer: So, you feel both prompts are clear, both are easy to understand, you have a lot to say about both, but this is better for a time limit.
>
> 10232010-03 (L1 Tagalog): Yeah, yeah.
>
> Interviewer: So, that's all really helpful and I understand you did not choose A because it's not so familiar to your personal background. But just this language here (looking at prompt A) was there anything confusing or not clear about prompt A?
>
> 10232010-04 (L1: Punjabi) No. I understood that statement but I didn't have enough ideas about that. So that is the reason I didn't choose.

At least two factors may help to explain the finding that interviewees did not report prompt length influenced their judgment of prompt difficulty or equivalence. One is that the maximum length of MELAB writing prompts is constrained by the test specifications. Although there is some variation in length, as demonstrated in the prompts described in Table 4.2, the difference in length is restricted by the test specs. The second factor is that the MELAB is a high-intermediate to advanced proficiency exam and the test takers are at a sufficiently high level of language proficiency to be relatively unaffected by minor differences in prompt length or linguistic difficulty.

Topic familiarity was highlighted as the single most important factor when choosing a prompt by the interviewees. Prompts in the personal domain were viewed as easier than other domains (public, occupational, or educational). However, prompt topics that interviewees had particular knowledge of; for example, because of their profession or studies were also favored because of their familiarity.

> 10232010-03 (L1 Arabic): Yes, I found myself familiar with the subject more and I have more vocabulary if I would like to write in this subject area. I found in my mind there is some resources, some examples, some ideas I can write. Better than other one. I can write the other one but I found myself be comfortable to write about the first topic. I have more resources in my mind.

Interviewees also said that prompts with several rhetorical cues were easier to respond to than those with a small number of such cues. The interviewees suggested that the rhetorical cues provide an outline of a response in terms of the structure expected and that the more cues there are, the easier it is to organize a response.

> 10232010-05 (L1 Arabic): With all this questions, because I'm, it's really memorable to me, like it changed my life, that's why even though they have lots of questions it help me build up my essay even better because if it will just, if it is just "what made it memorable and why" and didn't do other questions, I might not be able to put 250 words together, you know. So, for me, like hmmm, answering more questions is much better in creating an essay.

These findings from interviews with MELAB test takers resulted in prompt length being removed from the categorization of prompt characteristics but the number of rhetorical cues being retained. Although only seven test takers were interviewed, their comments on prompt length were unequivocal and made the removal of prompt length from the prompt categorization approach a straightforward decision. The interviewees' comments about topic familiarity also brought about a shift in how the prompts were categorized. The interviewees' strong feelings about topic familiarity indicated that the topic of the response was particularly important. The working category of expected person of response did not fully capture the comments made by the interviewees about topic familiarity. Topics that test takers felt were familiar may be related to their profession or studies and, hence topics that are favored by test takers may not be ones that simply elicit a first person response. That is, the expected person of the response may not be a sufficiently specific characteristic to operationalize the features of the response described by the interviewees. Prompt topics that are educational or occupational can be favored by test takers due to their familiarity.

Prompts are developed for the MELAB within four distinct domains; personal, public, occupational, and educational, following the guidance of the Common European Framework of Reference (Council of Europe, 2001). The interviewees indicated that personal domain prompts were generally viewed as easier than those situated in other domains but that further distinctions between personal and non-personal domain topics were important to capture within the prompt categorization framework. Hence, the expected person of the response was replaced in the prompt categorization framework by the domain the prompt is situated in.

As a result of the discussions with MELAB raters and the interviews with MELAB test takers, in addition to the initial categorization that was drawn from the literature, the three key characteristics of independent writing prompts that emerged were the domain, the response mode, and the number of rhetorical cues. Finally, influenced by the interviews with the MELAB test takers, a fourth prompt characteristic was added to the categorization. The final characteristic was the focus of the prompt (open or focused). This choice was influenced by two factors; first, interviewees consistently claimed that they selected prompts to respond to that offered them the opportunity to write a long response. Interviewees said this was possible when a prompt allowed them to write on a topic that they had existing knowledge of or a great familiarity with. Second, responses to prompt four (regrets) covered a very broad range of topics, much broader than the typical range of responses seen in MELAB essays. Many of the responses were memorable to raters as the writers described very personal and specific experiences from their lives. The range of topics covered in the responses to prompt four may, at least partially come about because of the broad scope presented to the writer by the prompt wording. Such prompts allow test takers to write about an event that is familiar to them and fits the description given during test taker interviews of a prompt that allows them to write a lengthy response because they have content readily retrievable from memory that can

be produced without a great deal of planning or organization. Hence, as this type of prompt was identified by both writers and raters as one that can elicit responses that differ distinctly from others, the distinguishing characteristic of prompt focus was added to the prompt characterization framework.

This category (focus) incorporates whether a prompt constrains a writer to produce a response on a very specific topic, as prompts one (work week) and two (banning foods) do (see p.57 for the prompt wording). This category also incorporates prompts where the writer is allowed a great deal of freedom to respond and still be considered on topic. This open to focused continuum is the fourth prompt characteristic and is tagged as prompt focus. This category is similar in definition to that used by Lim (2010) and named task constraint in Lim's work. As Lim (2010) also developed an approach to categorization based on MELAB writing prompts, this characteristic seems to be an identifiable characteristic of MELAB writing prompts and quite possibly in other testing programs that employ large numbers of independent writing prompts.

The four distinguishing characteristics, as described above and derived initially from the literature and a reading of authentic responses to several prompts, and then clarified with the input of raters and test takers are presented in Table 4.3.

Table 4.3: Distinguishing characteristics of writing prompts

Domain	Prompts are categorized into four domains (educational, occupational, public, or personal) to explain the general topic area.
Response mode	Prompts elicit either a narrative or argumentative response.
Number of rhetorical cues	Prompts contain different numbers of rhetorical cues (defined as an instruction or question within the prompt that the writer must respond to), ranging from one to six cues in this study.
Open or focused	Prompts classified as open require little or no background information or contextualization. They do not confine the write to a particular direction that the response must take. Prompts classified as focused need contextualization to provide test takers with background information before they can attempt a response. These prompts restrict the range of possible responses that the writer may take to compose a direct response to the prompt.

The following two prompts exemplify the four prompt characteristics and how they help to distinguish different prompts.

Sample Prompt 1:

> *In everyone's life, some days are particularly memorable. Describe what happened on one such day that you have experienced in your own life. What made it memorable and why? Include specific details such as where you were, who else was there, and so on, that will convince the reader how memorable this day was.*

Domain: Personal
Response Mode: Narrative
Number of rhetorical cues: 6
Open or Focused: Focused
In this prompt, the rhetorical cues are:

1. describe what happened on one such day that you have experienced in your own life
2. What made it memorable
3. and why
4. Include specific details such as where you were
5. who else was there
6. that will convince the reader how memorable this day was

Sample Prompt 2:

> *The use of computers in elementary school classrooms is becoming very widespread. Do you believe that this will be helpful for student learning? Support your opinion with reasons and specific examples.*

Domain: Educational
Response Mode: Argumentative
Number of rhetorical cues: 3
Open or Focused: Open

In this prompt, the rhetorical cues are:

1. Do you believe that this will be helpful for student learning?
2. Support your opinion with reasons
3. And specific examples

This chapter has presented a framework for how the writing prompts were categorized. After a multistage process that incorporated a review of relevant literature, a review of authentic writing prompts and their associated responses, and interviews with experienced raters and authentic test takers, four key prompt characteristics were arrived at.

- Domain
- Response mode
- Number of rhetorical cues
- Focus

Categorizing writing prompts in this way allows prompts to be studied systematically. According to the literature, experienced raters, and test takers these are the characteristics of writing prompts that most assist in distinguishing one independent writing prompt from another. Hence, these are likely to be relevant prompt characteristics to investigate when attempting to learn whether different writing prompts may have an effect on the written language elicited. Studying prompts that differ according to

these prompt characteristics maximizes the opportunity to detect prompt effect and to be able to attribute any prompt effect detected to a particular prompt characteristic or characteristics.

4.3 Selecting the prompts for the main study

The approach to writing prompt categorization described above allows for a systematic investigation of the relationship between different types of writing prompts and the written product elicited. Based on the prompt categorization framework described above, six retired MELAB writing prompts were selected for the main phase of this study. These prompts (see Table 4.4) were selected because they operationalize all of the prompt characteristics identified above (Table 4.3). Two of these six prompts (#4 and #5 in Table 4.4) were drawn from the prompts discussed in Section 4.2 above and four of the six prompts were newly selected for the main study.

The prompts were sourced from the MELAB database and were not developed exclusively for this study. Hence, there is some nesting of prompt characteristics in the prompts; that is the different prompt characteristics are not perfectly evenly distributed across the six prompts and each prompt does not operationalize a unique set of prompt characteristics. Working with existing prompts and responses, this limited degree of nesting was unavoidable. The advantage of working with authentic prompts and authentic test taker responses (see Section 2.5.2) was considered more important than writing new prompts to fit the prompt categorization framework. Working with existing prompts and responses also allowed for access to the MELAB database and the very diverse population of test takers. In summary, the collection of an authentic dataset was prioritized over the creation of a set of prompts that would have avoided nesting in the design.

Table 4.4: Prompts used in main study

Prompt #	Prompt ID	Keywords	Prompt
1	95	child psychologists	Some child psychologists believe that the peer groups children play with influence their character and personality development more than the children's parents do. The psychologists say children are more interested in fitting in with their friends than behaving the way their parents want them to. Do you agree or disagree with these psychologists? Explain your point of view.
2	214	government leaders	How important is it to know about the personal life (e.g. health, personal relationships, youthful mistakes) of government leaders? What things should be made public? What things should be kept private? Give your opinion and support it with reasons.
3	100	professions	Would you rather have the same profession all your life or change jobs often? Explain the reasons for your preference.
4	108	standard work week	In some countries such as the United States, the standard work week is 5 days, 8 hours a day. Some companies are considering changing their work week to 4 days, 10 hours a day. What are the advantages and disadvantages of the two options? Which do you think would be better for society? Give reasons to support your opinion.
5	115	memorable days	In everyone's life, some days are particularly memorable. Describe what happened on one such day that you have experienced in your own life. What made it memorable and why? Include specific details such as where you were, who else was there, and so on, that will convince the reader how memorable the day was.
6	73	Mistakes	It is often said that we learn more from our mistakes than our successes. Tell about a mistake that you once made and learned something from.

Table 4.5 shows how the six prompts in Table 4.5 are categorized.

Table 4.5: Categorization of study prompts

Prompt # (ID)	Keywords	Domain	Response mode	# of rhetorical cues	Focus
1 (95)	child psychologists	Public	Argumentative	2	Focused
2 (214)	government leaders	Public	Argumentative	5	Focused
3 (100)	professions	Occupational	Argumentative	2	Open
4 (108)	standard work week	Occupational	Argumentative	6	Focused
5 (115)	memorable days	Personal	Narrative	6	Focused
6 (73)	Mistakes	Personal	Narrative	2	Open

Section 4.4 will explain how the responses to the six prompts were collected to make up the operational dataset.

4.4 Collecting the dataset

60 responses to each of the six prompts were collected from past administrations of the MELAB. For each of the six writing prompts, the aim was to collect 20 responses at three distinct levels of language proficiency. This number of 20 responses at each proficiency level was driven by advice from a statistics consultant at the Center for Statistical Consultancy and Research (CSCAR) at The University of Michigan. A minimum sample size of 20, within each cell of the matrix that makes up the overall dataset was recommended. This number allows for an observation of trends at the narrowest level of interest (control for proficiency) within the dataset. Each response within the dataset was written by a different test taker; that is, there are 360 individual writers represented in the dataset.

The control of language proficiency helps to guard against any prompt effect that may be detected being attributable to subgroups within the sample population that may be at different levels of proficiency. For example, if the test takers who responded to one prompt happened to be at a particularly high level of proficiency, then the written products would likely reflect the test takers' proficiency level. An analysis of these responses may indicate that the prompt elicits complex or sophisticated language, when these observed features may be attributed to the population and not the prompt. Hence, controlling for language proficiency within each prompt-specific sub-group was necessary to focus the study on the effect of the prompt characteristics.

Language proficiency was controlled for by identifying test takers for inclusion in the sample population based on their score on the grammar, cloze, vocabulary, and reading (GCVR) section of the MELAB (see Table 4.1). GCVR scaled scores are reported on a 0-100 point scale. Raw scores are converted to scaled scores and the scores across test forms (a unique form of the MELAB is administered once a month) are

equated to ensure that scores are equivalent from form to form. The three levels of language proficiency that were identified for this dataset were as follows:

- Low proficiency band = GCVR 45-74
- Medium proficiency band = GCVR 75 – 84
- High proficiency band = GCVR 85-100

Table 4.6 illustrates the total number of responses within the dataset, broken down by prompt and by proficiency band.

Table 4.6: Distribution of responses by prompt and proficiency band

Prompt #	**# of responses in low-proficiency band**	**# of responses in medium-proficiency band**	**# of responses in high-proficiency band**	**Total # of responses**
1 (95)	20	20	20	60
2 (214)	20	20	20	60
3 (100)	20	20	20	60
4 (108)	20	20	20	60
5 (115)	20	20	20	60
6 (73)	20	20	20	60
				Dataset Total = 360

The sample population was not divided into sub-groups by scores on the writing section of the MELAB to avoid the risk of predetermining the writing features that could be found at each of the three proficiency levels. Using these scores to group responses to prompts into three bands would have predetermined, to some extent the type of textual features within the dataset. Using GCVR scores to group responses into distinct proficiency bands helps guard against this possibility. In addition, the GCVR section is the most reliable section of the MELAB (r= 0.93-0.95; CaMLA, 2013) and hence, provides a consistent measure of test takers' language proficiency.

The comparability of GCVR scores across the six writing prompts within each proficiency group was checked using ANOVA. The purpose of doing so was to check for significant differences in the score profiles awarded to individuals within each proficiency group by writing prompt. The results of the ANOVAs are shown in tables 4.7 to 4.9 below.

Table 4.7 shows the ANOVA results for the dependent variable GCVR score for the low proficiency group. The independent variable is the writing prompt.

Table 4.7: ANOVA results for high low proficiency group

	Sum of squares	df	Mean Square	F	Sig.
Between Groups	1109.5	5	221.9	5.5	.0001
Within Groups	4595.8	114	40.3		
Total	5705.3	119			

There was a significant effect of writing prompts on the GCVR scores of the low proficiency group $F(5, 114) = 5.5$, $p<.05$. These data indicate that there are significant differences in the GCVR scores awarded to test takers in the low proficiency group, depending on the prompt.

Table 4.8 shows the ANOVA results for the dependent variable GCVR score for the medium proficiency group. The independent variable is the writing prompt.

Table 4.8: ANOVA results for medium proficiency group

	Sum of squares	df	Mean Square	F	Sig.
Between Groups	80.1	5	16.0	1.82	.1147
Within Groups	1004.9	114	8.8		
Total	1084.9	119			

There was no significant effect of writing prompts on the GCVR scores of the medium proficiency group $F(5, 114) = 1.82$, $p>.05$.

Table 4.9 shows the ANOVA results for the dependent variable GCVR score for the high proficiency group. The independent variable is the writing prompt.

Table 4.9: ANOVA results for high proficiency group

	Sum of squares	df	Mean Square	F	Sig.
Between Groups	98.9	5	19.8	2.2	.0596
Within Groups	1027.6	114	9.0		
Total	1126.6	119			

There was no significant effect of writing prompts on the GCVR scores of the high proficiency group $F(5, 114) = 2.2$, $p>.05$.

These ANOVA results show that the GCVR scores did not differ significantly by writing prompt for the medium and high proficiency groups. However, that was not the case for the low proficiency group. For the low proficiency group, there were significant differences in GCVR scores by writing prompt. Descriptive statistics for the GCVR scores for each of the six writing prompts are shown in Table 4.10.

Table 4.10: Descriptive statistics for low proficiency group GCVR scores

Prompt	Mean GCVR score	SD	Minimum	Maximum
73	70.75	2.77	65	74
95	68.15	5.15	52	74
100	69.3	5.13	52	74
108	63.15	5.69	52	70
115	68.7	5.23	52	74
214	62.85	11.03	46	74

The descriptive statistics show that the GCVR score profiles for prompts 73 and 214 are quite different, with relatively large differences in range and standard deviations. During data collection it was challenging to identify MELAB essays from test takers who had GCVR scores that were either very low or very high. Indeed, finding responses to specific prompts with GCVR scores below 70 was not at all simple. MELAB Writing prompts are administered on only a limited number of occasions to prevent exposure and some prompts simply had few responses available and this was particularly problematic at the extreme ends of the scale. The aim of this study is to investigate the effect of specific prompts that most exemplify certain prompt characteristics and as a result, there was little option other than to use the responses available for prompts 73 and 214 for the low proficiency groups.

This limitation in the dataset means that it would likely be inadvisable to compare the effects of the writing prompts on responses by proficiency band, especially for the low proficiency group. Given the

significant differences in language proficiency in the low proficiency group, any differences in the written responses of these test takers (within this group) may be attributable to the language proficiency differences as to characteristics of the different prompts. This limitation will be considered when analyzing the effect of the prompts on written responses for test takers at specific proficiency levels. Beyond this limitation to analyzing data across proficiency bands, the significant differences in GCVR scores in the low proficiency group do not present any risk to the main aims of the quantitative study, namely to investigate the effect of different prompt characteristics on the textual features of responses for the MELAB test population as a whole.

Table 4.11 shows the results of the ANOVA with the writing prompt as the independent variable and writing score as the dependent variable.

Table 4.11: ANOVA results for writing score awarded

		Sum of Squares	df	Mean Square	F	Sig.
Writing score	Between Groups	235.356	5	47.071	1.149	.334
	Within Groups	14507.300	354	40.981		
	Total	14742.656	359			

There was no significant effect of writing prompt features on the original score awarded, $F(5, 354) = 1.149$, $p>.05$. These data indicate that the writing prompts studied in this work elicit responses that do not differ significantly in writing score awarded.

The correlation (Spearman's Rho) between Writing scores and GCVR scores is 0.72, indicating a moderate to strong relationship between the two sets of scores, indicating that GCVR scores are a good control for language proficiency but do not provide the same measurement information about the sample population as the Writing scores. That is, the findings of the ANOVA and correlation data indicate that GCVR scores are appropriate for use as a language proficiency control with the caveat that there are significant differences between GCVR scores by prompt for the low proficiency group in the sample population.

Using the criteria described above (60 responses per prompt at three distinct levels of language proficiency), 360 MELAB essays were collected. The responses were drawn broadly from test forms administered between 2006 and 2011. Multiple test centers in both Canada and the United States were represented in the sample, along with centers in Austria, Japan, and China.

Table 4.12: Number of test taker responses collected at specific MELAB Test Centers

Test Center	**Number of responses**	**Percentage**
Toronto	93	25.8
Vancouver	66	18.3
Calgary	45	12.5
Detroit	43	11.9
Ann Arbor	34	9.4
Shanghai	24	6.7
Tokyo	18	5
Others	37	10.3
TOTAL	360	100

4.4.1. The sample population

Of the 360 test takers in the sample, 54.6% were female and 45.4% were male. The average age of the test takers was 25. There were many native languages represented in the sample population. Table 4.13 shows the most represented native languages.

Table 4.13: Well-represented native language backgrounds

Language	**Number of Test Takers**	**Percentage**
Tagalog	112	31.1
Chinese	81	22.5
Arabic	55	15.3
Farsi	26	7.2
Malayalam	20	5.6
Japanese	18	5
Spanish	13	3.6
Punjabi	11	3.1
Others	24	6.7
TOTAL	360	100

In Chapter 2, the importance of a diverse sample population was highlighted (see 2.4.2). Of particular importance is the language background, educational background, and gender balance of the sample population. Populations that lack diversity in these key areas may skew the findings, as described in the literature review of test taker features in Chapter 2 (see 2.4.1). The sample population for this study is drawn from test centers in five different countries and represents multiple different native languages, with no single first language predominating in the sample. The sample population is also evenly balanced in terms of gender representation. Overall, the diversity of the sample population helps guard against any confounding variable that may arise as a result from an over-representation of a single sub-group.

4.5 Analyzing the written products

The primary aim of this thesis is to investigate the relationship between the characteristics of independent writing prompts and the written products elicited by different prompts in a high-stakes test environment. In order to observe differences in the written product elicited by the writing prompts, it is necessary to analyze the written products in a systematic way.

4.5.1 Identifying textual features and discourse measures

The review of literature (see Chapter 2) revealed that using the score awarded to the written product as the criterion for determining prompt effect is ineffective. A more insightful approach has been to apply a set of discourse measures to the written product with the aim of detecting differences in the underlying textual features of the responses. As reported in Section 2.2, a discourse analytic approach to investigating prompt effect proved more successful in detecting trends in written language than using the score awarded. Table 4.9 shows the textual features that have been analyzed in the literature described in Chapter 2 and the specific discourse measures that were selected to operationalize the textual features. These examples of textual features and discourse measures are all drawn from studies that detected significant differences in written product. That is, these are textual features and discourse measures that have been shown to be successful in work that aims to detect prompt effect.

Table 4.14: Summary of textual features and discourse measures

Study	**Textual features**	**Discourse measures**
Greenberg (1981)	Fluency Syntactic complexity	Mean number of T-units Clauses per T-unit Words per clause Words per essay
Hirokawa & Swales (1986)	Syntactic complexity Accuracy Lexical sophistication	Subordination First person pronoun incidence Number of total errors Number of morphological errors Number of syntactic errors Proportion of Graeco-Latin vocabulary
Zhang (1987)	Fluency Syntactic complexity	Number of words Average sentence length Number of clauses per sentence
Spaan (1990)	Fluency Syntactic sophistication Accuracy Lexical range and sophistication Rhetoric	Number or words
Way, Joiner, & Seaman (2000)	Fluency Syntactic complexity Accuracy	Number of words Mean length of T-units Percentage of correct T-units
Cumming et al. (2005)	Lexical and syntactic complexity Grammatical accuracy Argument structure Orientation to evidence Verbatim use of source text	Text length, word length, type-token ratio Number of words per T-unit, Number of clauses per T-unit Quality of argument structure
O'Loughlin &Wigglesworth (2007)	Task fulfillment Coherence and cohesion Vocabulary and sentence structure	Structure and organization of the body Conjunctive and referential cohesion Subordination Repetition of key words
Kuiken & Vedder (2008)	Accuracy Syntactic complexity Lexical variation	Total number or errors Total number of severe errors
Ong & Zhang (2010)	Fluency Lexical complexity	Mean number of words per minute Type-token ratio

4.5.1.1 The textual features

As can be seen from the summary presented in Table 4.9, a wide range of different textual features have been analyzed. An even wider variety of discourse measures have been employed to operationalize the textual features. Some of the most commonly analyzed textual features are:

- Fluency
- Syntactic complexity
- Accuracy
- Lexical variation/sophistication

These four textual features of second language writing have been commonly explored across many of the studies that have investigated prompt effect using a discourse analytic approach. These four textual features will also be analyzed in this study. An additional textual feature that will be used in this study is cohesion. Cohesion appears in some of the studies described above and it has been selected as relevant for this work because of a hypothesis developed while reading MELAB essays to identify key prompt characteristics, as described in Section 4.2.2.3.

While reading responses to MELAB prompts one hypothesis that developed was that narrative and argumentative responses differed in terms of cohesion and use of vocabulary. Narrative responses tended to be lexically richer, with a greater use of low-frequency descriptive language and connective devices were used less mechanically to build cohesion than in argumentative responses. Many argumentative responses were thought to take a standard five-paragraph essay approach, with a thesis statement in the opening paragraph, three body paragraphs that contained the supporting detail, and a conclusion in the final paragraph that restated the thesis. Typically, these responses used mechanical transition markers (first, second, finally, in conclusion, etc.) to link ideas together. Indeed, the MELAB scoring rubric identifies the mechanical use of transition markers as a feature of the responses that identifies a developing writer rather than a proficient one. Hence, cohesion was added to the list of textual features to be analyzed.

4.5.1.2 The discourse measures

To identify the possible range of discourse measures that could operationalize these textual features, several computer programs were trialed to learn about the range and types of measures that were available. The programs trialed are shown in Table 4.15.

Table 4.15: Computer programs trialed

Program	Tools	URL
AntConc	Corpus linguistics; including concordances, clusters/N-grams, and collocates	http://www.laurenceanthony.net/antconc_index.html
COCA	Analyzes texts by word frequency based on a monitor corpus of 450 million words;	http://www.wordandphrase.info/
BNC	Analyzes texts by word frequency based on a corpus of 100 million words	http://www.natcorp.ox.ac.uk/
Coh-metrix	Computational linguistics, including measures of cohesion, syntactic complexity, and text readability	http://tool.cohmetrix.com/
Lextutor	Concordancer, vocabulary profiler, N-grams	http://www.lextutor.ca/vp/eng/

After trialing, COCA was selected to run the lexical analyses, due to its advantages in terms of the size of the corpus (450 million words) and the fact that it is a monitor corpus, meaning it is regularly updated with large amounts of recent language data that reflects contemporary language use. COCA also generates word frequency data for every word analyzed in a text via its Word and Phrase tool (URL is given in table above), which produces focused insight into the lexical sophistication of each text.

Coh-metrix was selected to run non-lexical analyses (see table 4.11 below for details), such as measures of cohesion and syntactic complexity, as it reports a broad range of measures that capture multiple data points representative of the traits of cohesion and syntactic complexity. In addition, the original design principles of Coh-metrix were to identify measures of cohesion in written language and cohesion was identified as a textual feature for analysis in this study. The broad range of measures allows the researcher to select the ones that most directly align with the aims of the research study. For these reasons, Coh-metrix was used to analyze responses for cohesion and syntactic complexity.

Table 4.16 shows the discourse measures that were selected to operationalize the five key textual features initially chosen to operationalize the construct of second language writing proficiency.

Table 4.16: Initial selection of discourse measures

Major textual features	Representative discourse measures
Fluency	• Response length
Syntactic complexity	• Average paragraph length • Average sentence length • Number of words before main verb
Lexical sophistication	• Type-token ratio • Lexical density • Lexical frequency profile • Average word length
Cohesion	• Incidence of all connectives • Incidence of causative connectives • Incidence of logical connectives • Incidence of temporal connectives • Incidence of additive connectives
Accuracy	• Total number of errors • Number of errors per 100 words

Many of these textual features are captured and assessed in the MELAB Writing rating scale. While fluency is not directly addressed in the scale, all other textual features are addressed explicitly. The descriptors presented below, from score point 73, in the middle of the MELAB Writing rating scale (see Appendix 2 for full scale) help exemplify how the textual features in Table 4.11 are present within the scale.

> **73** Topic development is present, although limited by incompleteness, lack of clarity, or lack of focus. The topic may be treated as though it has only one dimension, or only one point of view is possible. In some 73 essays, both simple and complex syntactic structures are present, but with many errors; others have accurate syntax but are very restricted in the range of language attempted. Morphological control is inconsistent. Organization is partially controlled, while connection is often absent or unsuccessful. Vocabulary is sometimes inadequate, and sometimes inappropriately used. Spelling and punctuation errors are sometimes distracting.

References to syntactic complexity (both simple and complex syntactic structures are present) and lexical sophistication (Vocabulary is sometimes inadequate, and sometimes inappropriately used) are clear at this score point. Accuracy is also explicitly referenced in the scale (complex syntactic structures are present, but with many errors; others have accurate syntax but are very restricted in the range of language attempted. Morphological control is inconsistent.) Cohesion is also referenced in the scale via mention of overall organization and use of connectors (Organization is partially controlled, while connection is often absent or unsuccessful.)

The specific discourse measures that were identified to operationalize the textual features are described here.

1. Response length (word #)

Essay length is commonly used as a measure of fluency (Perkins, 1980; Chenoweth & Hayes, 2001). Development of ideas within a response is difficult to achieve without a certain number of words, and writing assessment research indicates that length of response is one of the strongest predictors of final score awarded.

2. Average paragraph length (APL)

Very short paragraphs are typically evidence of a lack of development in a certain part of the response. In MELAB responses, one- or two-sentence paragraphs as supporting details are not uncommon. Average paragraph length is one criterion measure that can be used to gain an insight into the fluency of the writer and the development of the response.

3. Average sentence length (ASL)

Longer sentences tend to require that the writer have command of more complex syntactical rules (Zhang, 1987). Therefore, this was one measure used to assess syntactic complexity. Average sentence length is a key measure in readability formulae, such as the Flesch Reading Ease Score and The Lexile Framework for Reading.

Run-on sentences can be a confounding feature for this discourse measure, however there was no prevalence of run-on sentences in the essays as they were transcribed. The broad representation of first languages in the dataset helps guard against particular traits of certain native-language backgrounds.

4. Syntactic left-embeddedness (SYNLE)

This measure was used as an indicator of syntactic complexity (Banerjee, Yan, Chapman, & Elliott; 2015). It counts the average number of words that occur before the first verb in a sentence. In simple independent clauses, the verb tends to occur relatively early in the sentence. A longer delay before the verb occurs is a likely indicator of a more advanced feature such as a dependent clause (Chapman, Collins, Dame, & Elliott; 2014).

5. Type/ token ratio (TTR)

The type-token ratio reflects the number of unique words in a text divided by the total number of words, and is reported as a decimal between 0 and 1. This variable measures the range of vocabulary in a text (Engber, 1995; Laufer & Nation, 1995). As the type/token ratio is sensitive to the length of the sample analyzed, a sample of 190-210 words from each essay was used.

6. Lexical density (LEXDENS)

Lexical density refers to the ratio of number of content words (verbs, nouns, adjectives) to the total number of words in a text (Halliday, 1985). This variable provides a measure of the information load in a text as the content words typically convey content within a text, as opposed to the function words (prepositions, pronouns, conjunctions, etc.).

7. Lexical frequency profile (FREQ)

Word frequencies were investigated using the Corpus of Contemporary America English (COCA), which is a monitor corpus encompassing 450 million words, as of 2012. COCA divides words into three frequency bands. The first (FREQ1) represents the 500 most common words of English. The third band (FREQ3) represents more infrequent words, beyond the 3000 most common words of English, and band two (FREQ2) contains are all words that fall between band 1 and band 3. COCA also classifies words by domain and reports a fourth band (FREQAC) whereby words are categorized as 'academic.' This occurs when, based on the texts that comprise the corpus, a word appears in academic texts (such as research journals) at least twice as often as in other types of texts or its overall frequency average. A word tagged as 'academic' is not necessarily a more difficult word, but it does reflect the mode of academic writing.

Because COCA often does not recognize a misspelled word, it tends to treat the word as highly infrequent (it defaults to band 3). For this reason, before running an essay through COCA, spelling errors were corrected.

8. Average word length (AWL)

Average word length provides a measure of lexical sophistication in that texts with words that are longer, on average are likely to be more sophisticated (Biber, 1992; Grant & Ginther, 2000).

9. All connectives incidence (CONi)

Coh-Metrix was also used to count connectives to provide insights into connection, organization and rhetorical control (Halliday & Hassan, 1976; McNamara, Max, Louwerse, & Graesser, 2002). The value reported is the number of connectives per 1000 words.

10. Causative connectives incidence (CONCAUSi)

Causative connectives are connective devices such as *because, so*, and *consequently*.

11. Logical connectives incidence (CONLOGi)

Logical connectives per 1,000 words were recorded to explore a more specific subset of connectors that are appropriate to argumentative writing. Logical connectives include variants of *and, or, not*, and *if.*

12. Temporal connectives incidence (CONTEMPi)

Temporal connectives are connective devices that focus on aspects of time, such as *after, first, later, when*, and so on.

13. Additive connectives incidence (CONADDi)

Additive connectives are connective devices such as *also* and *moreover*.

14. Error count (total number of errors) and errors per 100 words

A global measure of accuracy was applied rather than classifying types of linguistic errors or ranking the effects of inaccuracies the error count guidelines. The approach recommended by Polio

(1997:139) was followed. Please refer to Appendix 3 for an account of how the error counts were performed. To account for essay length the number of errors per 100 words was also recorded.

4.5.2 Recording the discourse measure data

The 360 responses were transcribed by the researcher and the transcriptions were spot checked (one in ten responses) by a team of experienced MELAB raters to ensure that the transcriptions were a faithful representation of the original essays. All errors made by the test taker were retained. Where the test taker had crossed out a word or stretch of text, this was not included in the transcription.

During the initial round of analyzing the responses using Coh-metrix and COCA, the first 20 responses were run through the two programs three times to investigate the stability of the output from the two programs. No inconsistencies were observed in the output for any of the responses at any time. However, some issues were detected with some low-proficiency responses using both COCA and Coh-metrix.

Some of the responses within the low-proficiency band had major errors with mechanics; most notably spelling and punctuation. These errors would occasionally prevent Coh-metrix from producing any output and cause the program to freeze. Responses with these types of mechanical errors were also problematic for COCA. To address this issue, before running any response through COCA or Coh-metrix, spelling errors were corrected to ensure that COCA classified the words in the appropriate frequency band and so Coh-metrix could identify the appropriate part of speech. If COCA cannot recognize a word it defaults to the lowest frequency band (beyond the 3,000 most frequent words). Especially for low-proficiency writers who use very little low-frequency vocabulary, this COCA default can seriously distort the interpretation of lexical sophistication. In some low-proficiency group responses (approximately 15 responses in total), spelling errors were sometimes so confusing that the originally intended word choice could not be guessed at, even within the context of the response. In these cases, the response was discarded from the dataset and a replacement was found for the low-proficiency band for the prompt in question.

4.5.3 Finalizing the selection of discourse measures

The textual features represented by the discourse measures are drawn from the literature described in Chapter 2 and summarized in Table 4.11, however the number of discourse measures selected to operationalize the textual features was quite high. The high number of measures raised the concern that the output from the analyses may be challenging to interpret. As a result of these concerns, the following discourse measures were omitted from the work:

- Average paragraph length
- Lexical density
- Incidence of additive connectives

The reasons for the removal of these discourse measures were as follows.

Average paragraph length

Average paragraph length is not included in many studies that evaluate the features of written text. Average sentence length and syntactic left embeddedness were seen as better indicators of the syntactic complexity of a piece of writing than average paragraph length. Numerous responses in the dataset consisted of a single, very long paragraph, making average paragraph length an unreliable indicator of syntactic complexity for this dataset. The total length of each response was also included in the dataset and with response length and average sentence length being recorded, average paragraph length was a variable that could be dispensed with.

Lexical density

Lexical density was one of several discourse measures initially identified to evaluate the lexical sophistication of a text. The lexical frequency profile (because of the size and facility of the COCA corpus) and average word length (because of common use in similar studies) were seen as essential indicators of this linguistic feature, so the variable to omit was a straight choice between lexical density and type-token ratio. When recording the lexical density data, it appeared to be quite unstable and unpredictable. The values would vary more than expected for texts elicited from a single prompt and from writers at similar levels of general language proficiency. The type-token ratio, in contrast appeared more stable and easier to interpret as a simple measure of the number of unique words in a text. As a result, the lexical density data were omitted from the analyses.

Incidence of additive connectives

The inclusion of cohesion in this work was very important as it was hypothesized that narrative and argumentative responses build cohesion using different types of connective devices. Although reluctant to remove any measures from the analysis of cohesion, when working with the corpus data (recording lexical frequency profile data) and while transcribing the original MELAB essays, it became clear that the word “and” was used repeatedly and few other additive connectives were employed, even by writers in the high proficiency band. A brief examination of a few responses that had high incidences of additive connectives confirmed that this was due to the overuse of “and.” This did not seem to be a feature that revealed anything insightful into how writers build cohesion. Hence, the incidence of additive connectives was removed from the analyses.

The final selection of 15 discourse measures employed and the textual features of the written responses that were operationalized are shown in Table 4.17.

Table 4.17: Final selection of discourse measures

Major textual features	Representative discourse measures
Fluency	• Response length
Syntactic complexity	• Average sentence length • Number of words before main verb
Lexical sophistication	• Type-token ratio • Lexical frequency profile • Average word length
Cohesion	• Incidence of all connectives • Incidence of causative connectives • Incidence of logical connectives • Incidence of temporal connectives
Accuracy	• Total number of errors • Number of errors per 100 words

These textual features and representative discourse measures were the ones used in the analysis of the 360 essays written in response to six authentic MELAB writing prompts (60 responses per prompt). The analyses of these responses, using the discourse measures outlined above, represent the main quantitative approach taken in this work. The discourse measure data provides an understanding of how key textual features may vary within responses to different writing prompts. Statistical analyses of differences in these discourse measures (reported in Chapter 5) indicate how responses to the prompts differ in relation to these key textual features. This approach addresses the second research question: How do prompt characteristics affect the test-takers' final written product?

4.6 Summary

This chapter has described how the independent writing prompts analyzed in this work were categorized, how responses to these prompts were identified for analysis, and how the responses were analyzed. The methodology presented in this chapter is intended to address research question 1 (what are the distinguishing characteristics of independent writing prompts?) and research question 2 (how do these characteristics affect the test takers' final written product?).

The first research question, regarding prompt categorization has been directly addressed in this chapter. After a triangulated approach to the prompt categorization process that involved the consideration of previously developed prompt taxonomies, an analysis of responses to different prompts, and input from authentic test takers and experienced raters, the prompts were categorized by four main characteristics:

- Domain
- Response mode
- Number of rhetorical cues
- Focus

The categorization of writing prompts into key characteristics allows for the possibility of understanding which ones are responsible for any differences in written product that may be detected. Without the prompts being categorized, it would be challenging to interpret any prompt effect (in terms of differences in written product) that may be identified. Having a better understanding of which characteristics may contribute to prompt effect will be helpful when attempting to make prompts more equivalent, in terms of the challenge and opportunity they present to test takers to demonstrate their writing proficiency.

The second research question will be addressed in the Chapter 5, which investigates the relationship between the characteristics of the prompt and the textual features of the response. It aims to uncover whether there is a significant and quantifiable relationship between different prompt characteristics and specific textual features of the response. This chapter (Chapter 4) set out how the textual features of the responses will be analyzed using a range of discourse measures (see Table 4.12). These discourse measures were selected to operationalize the textual features of fluency, syntactic complexity, lexical sophistication, cohesion, and accuracy, which all underlie the construct of second language writing proficiency.

Chapter 5 will report the results of the analyses of the 360 essays, using the discourse measures described above. Chapter 5 will also describe the statistical approaches taken to explore the discourse measure dataset. It is this quantitative exploration of the dataset that will provide the answer to the second research question and show how the responses varied depending on which prompt was addressed. These results should help provide a thorough understanding of how the prompt characteristics affected the textual features of the responses, and hence provide a measure of the prompt effect on a high-stakes ESL writing test.

Chapter 5 – Results of main study

Chapter 4 explained the framework for categorizing independent writing prompts, presented the prompts to be investigated, and the discourse measures that were used to analyze the essays elicited by the prompts in this study. The independent writing prompts from a high-stakes ESL exam have been categorized according to their key common characteristics. Based on this categorization, six writing prompts were selected for investigation and 60 essays, written in response to each of the six prompts were collected across three distinct proficiency bands. These essays were then analyzed by applying a range of discourse measures that focus on some key traits underlying the construct of second language writing proficiency.

This chapter reports the results of the quantitative investigation of whether the different prompt characteristics elicit written language that is similar or different in terms of the discourse measures. If the different prompt characteristics elicit written language that is consistent and stable in terms of the discourse measures, the prompt characteristics will not have a significant effect on the written product. This would be a good indication that the writing prompts are equivalent. However, if the written language elicited by the prompt characteristics differs significantly (based on the discourse measures), then the written products will vary based on the prompt characteristics and there may be a problem with prompt equivalence.

This chapter describes the quantitative dataset that has been collected and will report on the statistical procedures that were undertaken in order to answer the second research question with an emphasis on the written product.

> How do the writing prompt characteristics affect the test-takers' final written product?

As reported in the previous chapter, the prompt characteristics that are to be investigated are:

- Domain
- Response mode
- Number of rhetorical cues
- Focus

Based on the literature reviewed in Chapters 2 and 3, and the responses that were read when formulating the prompt categorization framework described in Chapter 4, the following hypotheses were formed about the possible relationships between prompt characteristics and the textual features of the written product.

> 1. Prompts that have the same characteristics will elicit written responses with no significant differences in textual features.
>
> 2. Prompts situated in the personal domain will elicit responses that differ significantly from prompts situated in other domains, according to some textual features.

3. Narrative responses will differ significantly from argumentative responses according to some textual features.

The textual features of the responses are operationalized, in this dataset as a range of discourse measures (see Table 4.16 on p.77 for the operationalization of textual features by discourse measures).

The following sections of this chapter report on the statistical properties of the dataset, with a focus on whether the discourse measure data is normally distributed. The importance of understanding the distribution and central tendency of the sample data is emphasized in many guides to statistics (c.f. Brown, 1988; Shavelson, 1995; Field, 2009). The distribution of these data also informs the choice of statistical methods that were applied to the dataset in order to answer the second research question. The remainder of the chapter provides a detailed account of the statistical methods applied and the results.

5.1 Descriptive statistics

Table 5.1 shows the descriptive statistics for the discourse measures that were applied to the 360 pieces of writing elicited by the 6 writing prompts. See p.73-75 for a description of all discourse measures.

Table 5.1: Descriptive statistics for discourse measures

	N	Range	Minimum	Maximum	Mean	Std. error	Std. deviation	Skewness	Kurtosis
Writing score	360	35	60	95	76.06	0.38	6.408	0.237	0.072
Word #	360	503	86	589	320.54	4.607	87.405	0.237	0.173
ASL	360	27.798	8.059	35.857	17.895	0.256	4.858	0.738	0.555
AWL	360	1.653	3.609	5.262	4.276	0.016	0.308	0.451	-0.21
SYNLE	360	13.166	1.278	14.444	4.259	0.084	1.599	1.458	4.915
CONi	360	125.34	49.08	174.419	100.05	1.086	20.61	0.189	-0.12
CONCAUSi	360	73.34	5.797	79.137	34.949	0.745	14.141	0.728	0.334
CONLOGi	360	89.893	17.677	107.57	51.437	0.935	17.735	0.592	0.039
CONTEMPi	360	58.824	0	58.824	18.48	0.531	10.076	0.87	1.372
TTR	360	0.63	0.37	1	0.533	0.004	0.068	1.688	11.285
FREQ1	360	26	65	91	80.17	0.237	4.49	-0.306	0.087
FREQ2	360	18	3	21	10.95	0.151	2.86	0.158	0.247
FREQ3	360	15	3	18	8.89	0.144	2.731	0.471	-0.137
FREQAC	360	21	0	21	2.27	0.169	3.208	0.997	1.878
TOTERR	360	94	1	95	25.6	0.752	14.277	1.133	2.072
ERR100	360	28.86	0.25	29.11	8.258	0.232	4.397	0.981	1.606

All of the discourse measures are continuous. For a reminder of the scales the discourse measures are reported on, please refer back to p.73-75. Table 5.1 (above) and Figures 5.1-5.4 (below) illustrate the general trends in the dataset.

5.2 Checking the assumptions of normality

The skewness and kurtosis statistics show that a majority of the discourse measures are normally distributed as their values are close to zero (Field, 2009). However, three of the fifteen appear to be non-normally distributed; SYNLE (syntactic left embeddedness), TTR (type-token ratio), and TOTERR (total number of errors). The SYNLE data is leptokurtic (kurtosis = 4.915) meaning that the distribution is peaked. This is also the case for the TTR data and also, but less so for the TOTERR data. To confirm that these discourse measures are non-normally distributed, p-p plots were created, as recommended by Field (2009, 139) to visually examine how closely the data fit to a normal distribution. The p-p plot plots "the cumulative probability of a variable against the cumulative probability of a particular distribution," (Field, 2009: 134). If the data are normally distributed they will fit along the diagonal of the plot that represents the normal distribution.

Figure 5.1 shows a p-p plot for a normally distributed discourse measure (CONi – incidence of all connectives). The tight fit of the variable to the diagonal of the normal distribution is clear, indicating that the CONi data is normally distributed. The p-p plots for SYNLE (Figure 5.2) and TOTERR (Figure 5.4) show that the discourse measure distribution differs from the normal distribution. However, the p-p plot for TTR (figure 5.3) does not suggest that the distribution is non-normal. The plotted line for TTR is quite parallel to that of the normal distribution indicating that the data are normally distributed.

Figure 5.1: p-p plot of data for incidence of all connectives

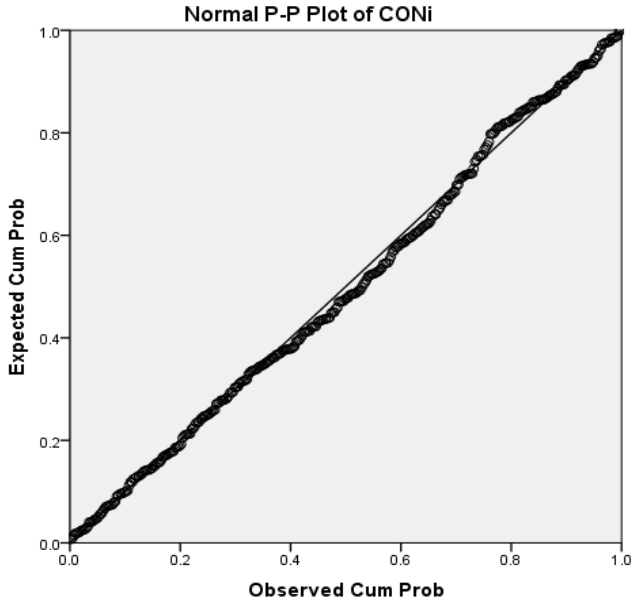


Figure 5.2: p-p plot of data for syntactic left-embeddedness

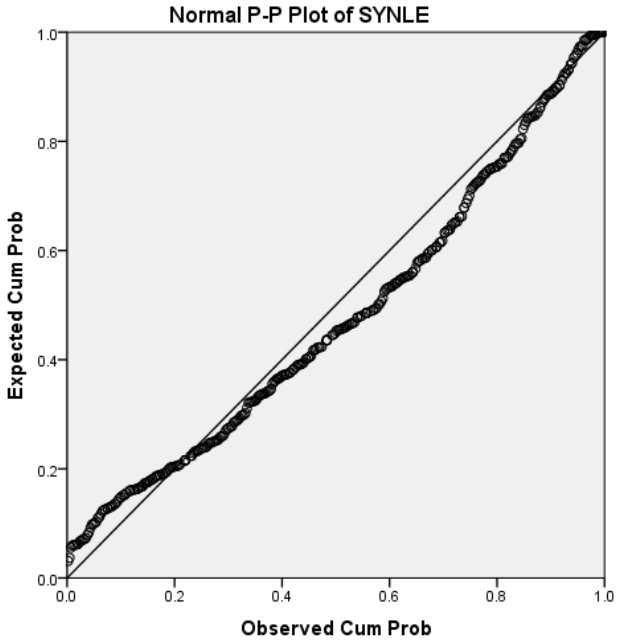


Figure 5.3: p-p plot of data for type-token ratio

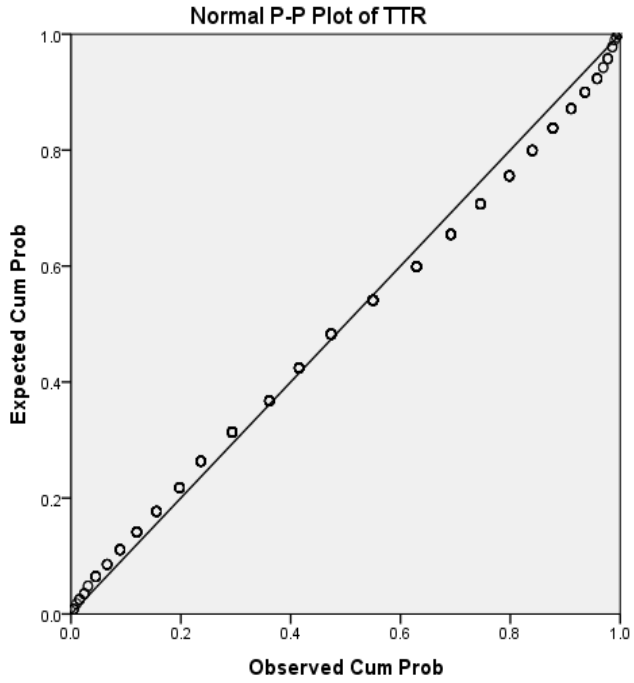
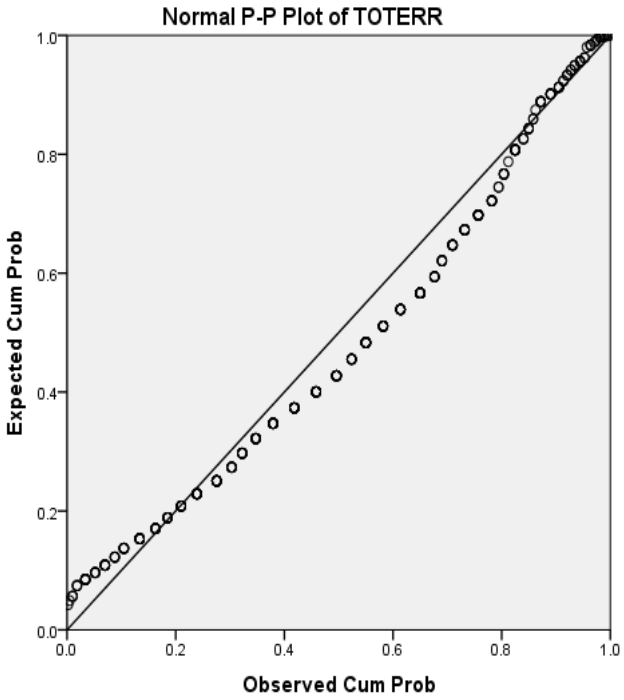


Figure 5.4: p-p plot of data for the total number of errors



The p–p plots indicate that the discourse measures SYNLE (syntactic left embeddedness), TTR (type-token ratio), and TOTERR (number of total errors) have non-normal distributions. This means that the non-normally distributed discourse measure data must be either transformed, or non-parametric tests must be run, or an approach that can deal robustly with violations of normalcy must be selected. The solution to this problem will be discussed in the sections that describe how the data set was handled for the:

- Factor analysis (see 5.3.1)
- Multivariate analysis of variance (MANOVA) (see 5.4.2)

Apart from the three discourse measures described above, the distributions for the other 12 discourse measures appear to be reasonably well centered and dispersed.

5.3 Preparing the dataset for analysis

The aim of the quantitative analyses of the discourse measure data is to investigate the written language elicited by the six prompts and to establish whether this language is consistent and stable across prompt types or whether there are significant differences in relation to one or more textual features under consideration. Responses to each of the six writing prompts were compared in relation to fifteen discourse measures used to represent key textual features (see Table 4.11).

The dataset comprises both multiple dependent variables (15 discourse measures) and independent variables (six writing prompts). The most appropriate statistical technique for examining group differences between multiple dependent and independent variables is multivariate analysis of variance (MANOVA). Performing MANOVA is preferable to running multiple analyses of variance (ANOVA) on each of the dependent variables and independent variables because MANOVA reduces the overall number of tests conducted on the dataset. Running many separate ANOVAs (one for each prompt's effect on a single writing prompt; 90 separate tests) greatly inflates the chance of making a Type I error. A Type I error would indicate that there was a significant mean difference between at least two of the discourse measures when no such difference actually exists. A further advantage of MANOVA over ANOVA, when there are multiple dependent variables is that MANOVA "by including all dependent variables, in the same analysis, takes account of the relationship between outcome variables." (Field, 2009: 586).

Although MANOVA reduces the likelihood of Type I error, compared to running multiple ANOVAs, the current dataset would still entail an analysis of 15 dependent (discourse) variables. This would still pose an inflated risk of Type I error in the results because of the large number of tests. One possible solution could be to remove some of the discourse measures from the analyses. However, the discourse measures selected were chosen specifically to operationalize important textual features (syntactic complexity, lexical sophistication, cohesion, accuracy, and fluency) that underlie the construct of second language writing proficiency. A solution was needed that would reduce the number of tests to be run without discarding discourse measures.

5.3.1 Reducing the number of dimensions for analyses

Factor analysis is an ideal tool for reducing "the number of dimensions necessary to describe the relationships among the variables," (Gardner, 2001: 242). The reduction in the number of factors, or hypothesized theoretical dimensions of second language writing proficiency will help to reveal whether these expected dimensions are reflected in the observed data. The term factor analysis refers to a group of statistical techniques that investigate "the relationships among a set of individual difference variables," (Gardner, 2001: 237). A common assumption of the factor analysis techniques is that "there are fewer components responsible for the relationships than there are variables," (Gardner, 2001: 237). The aim of performing a factor analysis on the current dataset is to reduce the number of latent variables (the discourse measures) that underlie the indicator variable of second language writing proficiency. The factor analysis will indicate how the latent variables relate to one another and will look for common themes among them, allowing some of them to be grouped together. This grouping of variables will assist with reducing the number of separate tests that need to be run to analyze the dataset.

It is important to note that the purpose of performing a factor analysis in this work is to explore the data and to analyze the factor structure (how the latent variables group together) of the sample. The factor analysis is not intended to yield findings that are generalizable beyond the sample collected to any broader population. That is, the aim of the factor analysis is descriptive rather than inferential (Field, 2009: 636).

Of the two main approaches to factor analysis, exploratory factor analysis and confirmatory factor analysis, an exploratory factor analysis was required. Exploratory factor analysis "begins with the relationships among the indicator variables and strives to uncover the dimensions underlying them," (Gardner, 2001: 238). Confirmatory factor analysis is used to test a model or theory that explains the nature of the relationship thought to exist between the latent variables. In the case of this study, the discourse measures had been grouped together along certain dimensions in previous studies (e.g. average word length and type-token ratio had been used to operationalize lexical sophistication), however some of the discourse measures used in this work differ from those used in previously published work on second language writing assessment (especially the use of COCA frequencies and certain Coh-metrix measures) so an exploratory rather than a confirmatory factor analysis was preferable.

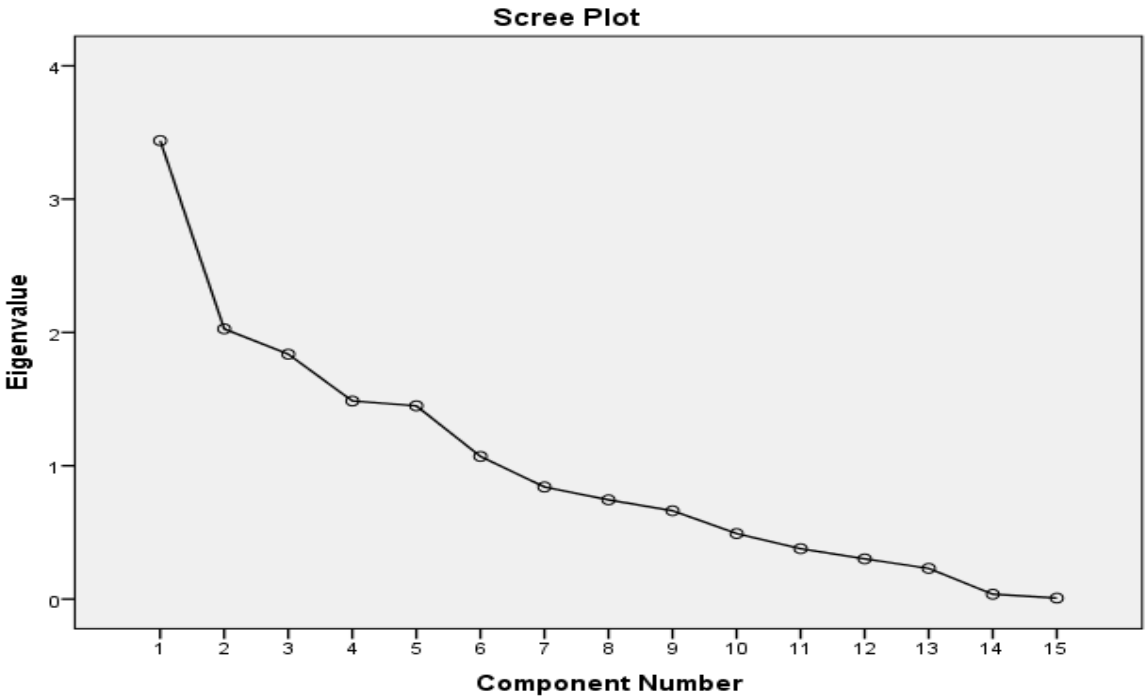
The term factor analysis is an umbrella term for several different approaches that have very similar aims; they attempt to reveal "the dimensionality underlying any given set of variables," (Gardner, 2001: 239). Using principal component analysis (PCA), all the variance associated with the variables is analyzed. With a relatively large number of variables for reduction and the appropriate choice of factor rotation, PCA offers a psychometrically sound and mathematically straightforward way of interpreting the relationships between the latent variables (Gardner, 2001: 242). Hence, principal component analysis was the approach adopted to reduce the number of latent variables and to gain an understanding of the relationships between the discourse measures.

The steps taken while running the factor analysis were guided by recommendations in the literature (Gardner, 2001; Field, 2009) and by advice provided to the author from the Center for Statistical Consultancy and Research at the University of Michigan. Field (2009) recommends first running the PCA without any rotation. The correlation matrix should first be observed to look for the extent to which the variables are correlated. This will influence the choice of factor rotation. The PCA should then be run again with the appropriate factor rotation. The desired factor solution will be identified based on the scree plot and on the eigenvalues of each of the factors (Gardner, 2001; Field, 2009). The following steps were followed when running the principal component analysis.

1. The PCA was first run without any rotation. This initial step was taken to allow for an examination of the correlation matrix between the latent variables. This large table is produced for reference in Appendix 4. The correlation matrix was reviewed for the number of correlations in excess of 0.6. The extent of collinearity within the correlation matrix provides some valuable information regarding the appropriate choice of factor rotation. As would be expected with the discourse measures selected to analyze the essays, some of them are highly correlated. Several of the measures that operationalize lexical sophistication, cohesion, and accuracy have high correlations (>0.6) with other measures that operationalize the same construct. However, as can be seen from the correlation matrix, there are no correlations in excess of 0.8, indicating no measures need to be removed from the analyses (Field, 2009: 648). As a result of this collinearity, an oblique rotation was selected, with the oblimin factor rotation being the most appropriate choice because of the size of the dataset (Field, 2009: 648).

2. The PCA was run again, this time with an oblimin rotation. In order to decide on a factor solution both the scree plot of the factors and the eigenvalues of the factors were observed (Gardner, 2001; Field, 2009). Figure 5.5 below shows the scree plot and Table 5.2 shows the eigenvalues associated with each factor.

Figure 5.5: Scree plot



The scree plot helps reveal the relative importance of each of the factors. Typically, the factors with high eigenvalues are shown in the steep descent of the curve. Many scree plots then tail off as the eigenvalues fall. The point at which the curve flattens is an indication of the number of factors likely to be in a factor solution.

The scree plot of this dataset is not particularly helpful when it comes to interpreting a factor solution. There is a leveling of the slope at factor 1, but after that the slope descends at a relatively steady angle until factor 14. It is not possible to make an interpretation regarding the likely number of factors in an optimal factor solution from this scree plot.

Even when the scree plot provides a clear indication of a possible factor solution, the final factor solution should not be determined just from a scree plot. It is also necessary to consider the eigenvalues for the factors. Table 5.2 shows the eigenvalues for all factors in the current dataset.

Table 5.2: Eigenvalues for all factors

Component	Initial Eigenvalues			Rotation Sums of Squared Loadings
	Total	% of Variance	Cumulative %	Total
1	3.439	22.924	22.924	2.210
2	2.027	13.510	36.435	2.320
3	1.837	12.247	48.682	1.908
4	1.486	9.907	58.588	2.035
5	1.449	9.663	68.252	1.706
6	1.070	7.134	75.386	1.191
7	.841	5.606	80.992	1.308
8	.744	4.962	85.953	2.184
9	.662	4.416	90.370	
10	.491	3.276	93.645	
11	.378	2.518	96.163	
12	.302	2.011	98.175	
13	.230	1.533	99.708	
14	.037	.244	99.952	
15	.007	.048	100.000	

Extraction Method: Principal Component Analysis.

Recommendations regarding the eigenvalue that indicates a factor should be retained vary from 1.0 (Kaiser, 1960) to 0.7 (Joliffe, 1986). As Table 5.2 shows, there are six factors with an eigenvalue greater than 1.0. These six factors account for 75.39% of total variance. There are eight factors with eigenvalues greater than 0.7 and these eight factors accounts for 85.95% of total variance. It is therefore likely that the optimal factor solution will have between six and eight factors.

After looking at both the scree plot and the eigenvalues, the PCA was run again several more times with oblimin rotation. The aim was to find an optimal factor solution with as few double loadings as possible and to look for a solution that was understandable theoretically, based on models of second language writing proficiency. After running the PCA with factor solutions from four to 11 factors, the optimal solution was one with seven factors as shown in Table 5.3 below.

Table 5.3: Factor analysis pattern matrix (see Table 4.11 for list of discourse measures)

Pattern Matrix

	Component						
	1	2	3	4	5	6	7
FREQ 1	.996						
FREQ 3	-.805						
FREQ 2	-.783						
TTR	-.358						
CONCAUSi		.847					
CONLOGi		.841					
CONi		.782					.414
ERR/100			.978				
TOTERR			.932				
AWL				.919			
FREQ AC				.914			
SYNLE					.863		
ASL					.863		
word #						.991	
CONTEMPi							.978

Extraction Method: Principal Component Analysis.

Rotation Method: Oblimin with Kaiser Normalization.

As can be seen in Table 5.3, the principal component analysis solved with a seven-factor solution. Although several different solutions were considered, the seven-factor solution had the advantages of including only one double factor loading (CONi – incidence of all connectives) and, more importantly that the solution was theoretically sound. The grouping of the seven factors along theoretical lines is addressed below with a more detailed discussion thereafter:

- Factor 1 (FREQ1, FREQ 2, FREQ 3 – lexical frequency profile, and TTR – type-token ratio) – all four discourse measures are representations of lexical sophistication.
- Factor 2 (CONCAUSi – incidence of causative connectives, CONLOGi – incidence of logical connectives, CONi – incidence of all connectives) – all three discourse measures are representations of cohesion.
- Factor 3 (ERR/100 – number of errors per 100 words, TOTERR – total number of errors) – both discourse measures are representations of accuracy.
- Factor 4 (AWL – average word length, FREQ AC – academic vocabulary) – it was not anticipated that these two discourse measures would load together. However, they are both lexical measures and the words classified as academic in COCA (FREQ AC) are typically longer than non-academic words. Hence, this factor may be seen as either a grouping of words typically used in academic contexts or words that are longer than other words on average.

- Factor 5 (SYNLE – syntactic left embeddedness, ASL – average sentence length) – these two discourse measures are both representations of syntactic complexity.
- Factor 6 (word #) – is a measure of fluency.
- Factor 7 (CONTEMPi – incidence of temporal connectives) – is also a measure of cohesion but does not load with the other cohesion measures, most likely due to the differences between the narrative and argumentative responses within the dataset.

It should be remembered that the purpose of the factor analysis in this study is exploratory, with the aim of understanding only the relationships between the measures within the collected sample of data. There is no claim that the factor loadings are generalizable to any model of writing beyond the current sample.

5.3.2 The theoretical dimensions of the seven factors

The first factor groups four measures that represent lexical sophistication. The fact that FREQ1 has a positive value and the other measures have negative values is to be expected. The FREQ1 data from COCA is a measure of the percentage of a text that is made up of the 500 most common words in the COCA corpus (of 450 million words). That is, the FREQ1 measure represents words that are very common. FREQ2 and FREQ3 represent the percentage of a text that is made up of the less frequent words of English, including those words that are outside of the most common 3,000 words of English in the COCA corpus. Hence, the FREQ2 and FREQ3 data is a measure of lexical sophistication and the opposite of what the FREQ1 data represents. TTR (type-token ratio) is a measure of the degree of repetition of vocabulary within a text. The higher the TTR value, the less repetition of vocabulary there is in a text and hence, the more lexically sophisticated the text is. Therefore, the mix of positive and negative values within the first factor is to be expected.

The second factor groups three measures that all operationalize cohesion within a text. The three measures are the incidence of all connectives within a text (CONi), the incidence of logical connectives within a text (CONLOGi) and the incidence of causative connectives within a text (CONCAUSi). It is not surprising that these measures load together as they represent similar textual features. What is more surprising is that the incidence of temporal connectives, CONTEMPi does not load with the other measures of cohesion. The finding supports they hypothesis that different response modes (narrative and argumentative) may utilize different aspects of cohesion, possibly using different types of connectives, as mentioned in Section 4.2.2.

The third factor groups together two measures (TOTERR and ERR/100) that both represent accuracy. It would be surprising if these two measures had not loaded together as one is derived from the other. ERR/100 is an error count that is standardized for the length of the text.

The fourth factor is the one that is the most difficult to interpret and to classify. The two discourse measures both operationalize aspects of lexical sophistication; AWL is the average number of letters per word in a response and FREQ AC is the percentage of a text that is made up of words tagged as academic within the COCA corpus. Either of these measures would indicate that the word occurs substantially more often in an academic context than it does in any other context.

While this factor may be seen as a measure of academic language use within a text it could also be seen as a measure of the use of relatively long words within a text. The fact that these words from the COCA corpus load on a different factor from those words from the corpus not classified as academic, suggests that there is something identifying these words as distinct from other vocabulary in the responses. While average word length is clearly a feature of the academic categorization within COCA, the way that this academic categorization is determined in the COCA corpus (words that are prevalent in academic contexts over other contexts) makes it likely that this factor is an indicator of academic language use. This factor, while not completely unambiguous provides a further measure of lexical sophistication that loads separately from that of the first factor.

The fifth factor groups two measures that operationalize syntactic complexity. The two measures (SYNLE and ASL) represent different aspects of sentence complexity. Syntactic left embeddedness (SYNLE) records the average number of words that occur in a sentence before the first verb. Typically, sentences with more words before the first verb will be complex and are likely to demonstrate embedded or complex clause structure. The average sentence length (ASL) within a text provides a separate indication of syntactic complexity. The longer a sentence is, on average the more syntactically complex it is likely to be.

For second language writers, the occurrence of run-on sentences can distort average sentence length data as run-on sentences, while very long may not be syntactically complex and may be challenging to process for meaning. Given the diverse first language makeup of the sample population and the fact that the population is also at a relatively high level of proficiency, there is not an evident problem with run-on sentences in the dataset. As the essays were transcribed, run-on sentences were a linguistic features that were looked for but there was no evidence of a consistent pattern of run-on sentence incidence. Both of these measures (SYNLE and ASL) may be expected to group together and are representative of syntactic complexity.

The last two factors do not group with any of the others within the dataset. Factor 6 (word #) is a measure of fluency, or the length of response that a writer is able to produce. It was of interest that average sentence length (ASL) did not group with the length of the response. This is an indication that average sentence length is an indicator of syntactic complexity and that the length of the total response is a distinct latent variable. Finally, the incidence of temporal connectives (CONTEMPi) is the seventh factor in this factor solution. The potential reasons for this measure to group separately from the other measures of cohesion have been presented above.

In summary, the principal component analysis gives an insight into how the discourse measures used to analyze the essays group together. These groupings, or factor loadings help shed light on which textual features are being operationalized with the discourse measures. These are:

- Lexical sophistication
- Cohesion
- Accuracy
- Use of academic vocabulary

- Syntactic complexity
- Fluency

The principal component analysis presented a way forward for the analyses to be performed so as to reduce the number of separate MANOVAs that need to be run on the dataset, thereby reducing the risk of Type I errors.

5.4 Results of MANOVA and ANOVA analyses

During the factor analysis, when the seven-factor solution was identified, factor scores for each of the factors were retained as variables to be used in the MANOVA. Factor scores take the various different measures within each factor (for example; the total number of errors in a response and the number of errors per 100 words in factor 3 that represents accuracy). The factor score provides a single measure for each of the factors (a composite score). Simply explained, they represent a weighted average of the separate variables within a single factor. These factor score coefficients are calculated by taking the factor loadings and adjusting them "to take account of the initial correlations between variables," (Field, 2009: 634), an approach that stabilizes differences between units of measurement of variables within a factor. The factor scores in this thesis were calculated using the Bartlett method, which produces "scores that are unbiased and correlate only with their own factor," (Field, 2009: 635). The use of factor scores reduces the number of subsequent analyses to be run on the dataset.

5.4.1 Normality of the factor score distributions

Table 5.4 shows the descriptive statistics for the seven factor scores. In contrast to the descriptive statistics reported at the start of this chapter (on p.80), the skewness and kurtosis data indicate that the factors scores are better centered and dispersed than the raw data was. See p.88-89 for a description of the seven factors.

Table 5.4: Descriptive statistics for factor scores

	N	Range	Minimum	Maximum	SD	Skewness	Kurtosis
BART factor score 1	360	6.14	-3.67	2.47	1	-0.29	.18
BART factor score 2	360	5.28	-2.14	3.14	1	0.46	-.11
BART factor score 3	360	6.33	-2.19	4.14	1	0.87	1.19
BART factor score 4	360	5.79	-2.09	3.7	1	0.65	.36
BART factor score 5	360	6.72	-2.19	4.53	1	0.8	1.09
BART factor score 6	360	5.77	-2.92	2.84	1	0.07	.22
BART factor score 7	360	6.9	-2.41	4.48	1	0.63	1.34

Table 5.5 reports the results of Levene's Test, used to check assumptions of univariate normality for each of the dependent variables. It is necessary to check these assumptions before determining which MANOVA test statistic to adopt (Field, 2009: 604). The findings show that five of the seven factor scores meet the assumptions of univariate normality ($p>0.05$) but that two of the factor scores (4 and 7) do not ($p<0.05$).

Table 5.5 Levene's test of equality of error variances

	F	df1	df2	Sig.
BART factor score 1	2.213	5	354	.053
BART factor score 2	1.405	5	354	.222
BART factor score 3	.524	5	354	.758
BART factor score 4	4.987	5	354	.000
BART factor score 5	1.004	5	354	.415
BART factor score 6	1.536	5	354	.178
BART factor score 7	3.344	5	354	.006

As a result of the violations of univariate normality, the Pillai-Bartlett trace test statistic was selected to assess statistical significance of any differences observed. The Pillai-Bartlett trace test statistic was chosen because it is the most robust of the MANOVA test statistics to violations of assumptions of

normality, when sample sizes are equal (Field, 2009: 604). Gardner (2001: 272) adds that "if sample sizes are equal, violation of the assumptions of equivalence of the variance/covariance matrices has little effect, particularly on Pillai's Trace." Sample sizes are equal within this dataset (60 essays per prompt, or independent variable).

5.4.2 Running the MANOVAs

For the MANOVA analyses, the six writing prompts were the independent variables and the seven factor scores were the dependent variables. The purpose of the MANOVA is to investigate the relationships between the independent variables (the writing prompts) and the dependent variables (the seven factors, or dimensions that underlie the construct of second language writing proficiency). The results of the MANOVA will indicate whether there are significant differences in the factor scores depending on the differences in the prompt characteristics between the six writing prompts, or independent variables.

5.4.3 MANOVA results for all factors scores

Table 5.6 shows the results of the MANOVA analysis of the full dataset

Table 5.6: MANOVA run on full dataset

Effect	Value	F	Hypothesis df	Error df	Sig
Prompt	.998	12.541	35.000	1760.000	.000

Using Pillai Bartlett's trace, there was a significant effect of writing prompt characteristics on the factor scores, which broadly represent the construct of second language writing proficiency.

Table 5.7 shows the ANOVA summary table for the dependent variables (the seven factor scores). The writing prompts are the independent variables and each of the seven factor scores is the dependent variable for each ANOVA. The ANOVA summary table provides further insight into the relationships between the independent and dependent variables, indicating which of the seven factors contribute to the significant effect of the prompt characteristics revealed by the MANOVA.

Table 5.7: ANOVA summary table

Tests of Between-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Prompt	Factor score 1	56.678	5	11.336	13.273	.000	.158
	Factor score 2	29.404	5	5.881	6.316	.000	.082
	Factor score 3	4.769	5	.954	.953	.447	.013
	Factor score 4	131.665	5	26.333	41.005	.000	.367
	Factor score 5	51.580	5	10.316	11.879	.000	.144
	Factor score 6	18.822	5	3.764	3.917	.002	.052
	Factor score 7	31.412	5	6.282	6.789	.000	.087

There was a significant effect of writing prompt characteristics on all factor scores apart from factor 3 (accuracy). That is, there was a significant effect of writing prompt characteristics on the following factors:

- Lexical sophistication (factor 1)
- Cohesion (factor 2)
- Academic vocabulary use (factor 4)
- Syntactic complexity (factor 5)
- Fluency (factor 6)
- Incidence of temporal connectives (factor 7)

The partial eta squared values are an estimate of the effect sizes of each of the factors. Factor 4 (academic language use) has the largest effect size (.367) of the factors. Factor 1 (lexical sophistication) and factor 5 (syntactic complexity) also have modest effect sizes (.158 and .144, respectively) making these factors the ones of most interest for further analyses. The other effects sizes are small, indicating that these factors explain little of the total variance.

5.4.4 Discriminant function analysis: better understanding the dataset

In order to verify that the effect sizes indicated in the ANOVA results were appropriate for determining which factors to focus on in further (post-hoc) analyses, a discriminant function analysis was performed. Discriminant function analysis is strongly recommended by Field (2009: 615) as the best way to achieve a thorough understanding of a dataset containing multiple independent and dependent variables. The discriminant function analysis "finds the linear combinations of the dependent variables that best separates (or discriminates) the groups" (Field, 2009: 604).

Table 5.8 shows the initial output from the discriminant function analysis and the principal variates (the variance of each dependent variable) that account for all the variance in the model. The canonical correlation values are an indication of effect sizes. The table shows that the first two variates account

for 94.3% of variance and that both variates have large (.803) or moderately large (.486) effect sizes. The third variate also has a moderately large effect size (.317) but as it contributes only 5% of the variance, it is of much less interest than the first two variates.

Table 5.8: Discriminant function analysis

Function	Eigenvalue	% of Variance	Cumulative %	Canonical correlation
1	1.815	80.6	80.6	.803
2	.310	13.8	94.3	.486
3	.112	5.0	99.3	.317
4	.014	.6	99.9	.116
5	.003	.1	100.0	.052

Table 5.9 shows the standardized canonical discriminant function coefficients for the variates. Field (2009: 619) describes these as the most important part of the discriminant function output for interpreting the MANOVA findings. These coefficients in Table 5.9 are "the standardized versions of the values in the eigenvectors" (Field, 2009: 619). The table above suggests that only the first two functions should be considered when looking for factor scores that most contribute to variance. The canonical discriminant function coefficients (below) can be interpreted in the sense that "the ones with high correlations contribute most to group separation" (Field, 2009: 619). These variates will be the ones that are used in further rounds of analyses to learn more about the factors that contribute to the significant effects identified in the initial round of MANOVA.

Table 5.9: Standardized canonical discriminant function coefficients for the variates

	Function				
	1	2	3	4	5
Factor score 1 (lexical sophistication)	1.050	-.030	-.461	.017	.216
Factor score 2 (cohesion)	.090	.388	-.356	-.285	-.105
Factor score 3 (accuracy)	-.035	-.079	.336	.783	-.136
Factor score 4 (academic vocabulary)	1.224	-.381	.181	-.015	.163
Factor score 5 (syntactic complexity)	.150	.795	.188	.412	.207
Factor score 6 (fluency)	-.158	.121	.567	-.490	.517
Factor score 7 (temporal connectives)	-.216	-.484	-.430	.256	.750

Looking at function 1, factor score 1 (lexical sophistication) and factor 4 (academic vocabulary use) stand out as having markedly high coefficients. While function 2 had a smaller effect size and accounted for only 13.8% of variance, the coefficient of .795 for factor score 5 (syntactic complexity) is also of interest for further investigation.

It should be noted that both the discriminant function analysis and the ANOVA summary table from the initial MANOVA (Table 5.7) indicate that factors 1, 4, and 5 contribute most to the variance. As both of the standard approaches (Field, 2009; Gardner, 2001) to following up a MANOVA finding with a significant effect reveal that the same factors are contributing to variance, it is reasonable to proceed to look further at these three factors; lexical sophistication, academic vocabulary use, and syntactic complexity. Investigating whether there are significant differences in each of these three discourse measures will greatly aid in establishing how stable and comparable (or otherwise) the written language is that is elicited by the six writing prompts.

5.5 Select ANOVAs: a closer examination of significant differences in written product

An ANOVA was run on each of the three factors identified above and the results are presented below. The independent variable was the writing prompt and the dependent variable was the factor score.

Table 5.10 shows the ANOVA results for the dependent variable factor 1 that operationalizes the writing sub-skill of lexical sophistication.

Table 5.10: ANOVA results for factor 1 (lexical sophistication)

	Sum of squares	df	Mean Square	F	Sig.
Between Groups	56.678	5	11.336	13.273	.000
Within Groups	302.322	354	.854		
Total	359.000	359			

There was a significant effect of writing prompt characteristics on the factor that operationalizes lexical sophistication $F(5, 354) = 13.273$, $p<.05$. These data indicate that the writing prompts elicit responses that are not equivalent in terms of the degree of lexical sophistication produced in response to different prompts.

Table 5.11 shows the ANOVA results for the dependent variable factor 4 that operationalizes the writing sub-skill of academic vocabulary use.

Table 5.11: ANOVA results for factor 4 (academic vocabulary use)

	Sum of squares	df	Mean Square	F	Sig.
Between Groups	131.655	5	26.333	41.005	.000
Within Groups	227.335	354	.642		
Total	359.000	359			

There was a significant effect of writing prompt characteristics on the factor that operationalizes academic vocabulary use $F(5, 354) = 41.005$, $p<.05$. These data indicate that the writing prompts studied in this work elicit responses that are not equivalent in terms of the proportion of academic vocabulary used in response to different writing prompts.

Table 5.12 shows the ANOVA results for the dependent variable factor 5 that operationalizes the writing sub-skill of syntactic complexity.

Table 5.12: ANOVA results for factor 5 (syntactic complexity)

	Sum of squares	df	Mean Square	F	Sig.
Between Groups	51.580	5	10.316	11.879	.000
Within Groups	307.420	354	.868		
Total	359.000	359			

There was a significant effect of writing prompt characteristics on the factor that operationalizes syntactic complexity $F(5, 354) = 11.879$, $p<.05$. These data indicate that the writing prompts studied in this work elicit responses that are not equivalent in terms of the syntactic complexity produced in response to different writing prompts.

5.6 Post-hoc tests

After identifying between-group differences, as in the findings reported above, Field (2009) recommends post-hoc tests to further explore which groups (prompts, in this case) are responsible for the differences detected in both the MANOVA and ANOVA analyses. The post-hoc tests provide a further level of insight and help explore which prompts elicit responses with significant differences in the factors reported above. There are many post-hoc tests available to follow up a significant ANOVA

finding. The Bonferroni correction was used in this work as it guarantees control over Type I error (Field, 2009: 374). It does introduce the possibility of Type II error (the probability of rejecting an effect that does actually exist), but the greater concern here is avoiding Type I error. Falsely identifying differences between factors that do not exist may lead to inaccurate conclusions being drawn about the features of writing prompts that contribute to prompt effect. In this case, Type I errors may lead to the incorrect conclusion that certain prompt characteristics contribute to differential test performance when that is not the case. The aim of this thesis is to investigate the numerous prompt characteristics that clearly have a measurable effect on written responses and not to mistakenly explore effects that are the result of statistical error.

Table 5.13 shows the findings of the post-hoc test (Bonferroni) for factor score 1 (lexical sophistication). The data show that there are significant differences in lexical sophistication between the following prompts for factor score 1 ($p<.5$).

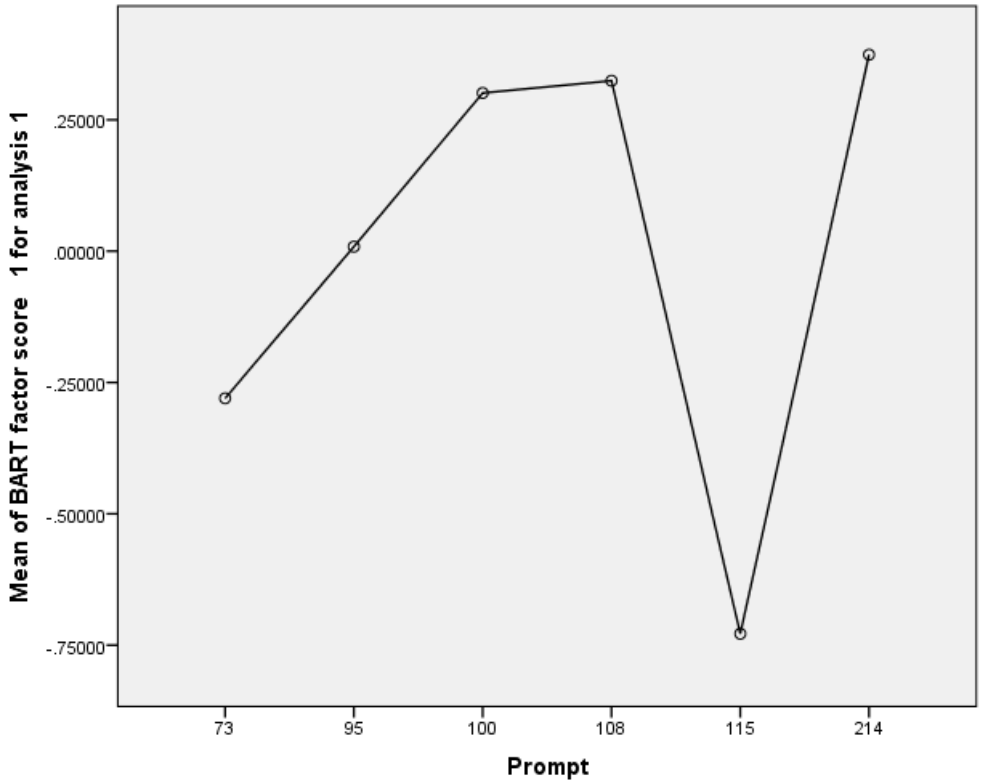
- 73 (mistakes) > 100 (professions)
- 73 (mistakes) > 108 (standard work week)
- 73 (mistakes) > 214 (government leaders)
- 95 (child psychologists) < 115 (memorable days)
- 100 (professions) < 115 (memorable days)
- 108 (standard work week) < 115 (memorable days)
- 115 (memorable days) > 214 (government leaders)

Table 5.13: Post-hoc test for factor score 1 (lexical sophistication)

Prompt			Mean Difference	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Bonferroni	73	95	-.290	0.17	1.00	-.79	.21
		100	-.581*	0.17	.01	-1.08	-.08
		108	-.604*	0.17	.01	-1.10	-0.11
		115	.450	0.17	.12	-.05	.95
		214	-.654*	0.17	.00	-1.15	-.16
	95	73	.290	0.17	1.00	-.21	.79
		100	-.290	0.17	1.00	-.79	.21
		108	-.320	0.17	.93	-.81	.18
		115	.737*	0.17	.00	.24	1.24
		214	-.370	0.17	.46	-.86	.13
	100	73	.581*	0.17	.01	.08	1.08
		95	.290	0.17	1.00	-.21	.79
		108	-.020	0.17	1.00	-.52	.48
		115	1.029*	0.17	.00	.53	1.53
		214	-.070	0.17	1.00	-.57	.43
	108	73	.604*	0.17	.01	.11	1.10
		95	.320	0.17	.93	-.18	.81
		100	.020	0.17	1.00	-.48	.52
		115	1.053*	0.17	.00	.55	1.55
		214	-.050	0.17	1.00	-.55	.45
	115	73	-.450	0.17	.12	-.95	.05
		95	-.737*	0.17	.00	-1.24	-.24
		100	1.029*	0.17	.00	-1.53	-.53
		108	1.053*	0.17	.00	-1.55	-.55
		214	-.050*	0.17	.00	-1.60	-.60
	214	73	.654*	0.17	.00	.16	1.15
		95	.370	0.17	.46	-.13	.86
		100	.070	0.17	1.00	-.43	.57
		108	.050	0.17	1.00	-.45	.55
		115	1.102*	0.17	.00	.60	1.60
*. The mean difference is significant at the .05 level.							

Figure 5.6 presents the means plots of the differences in lexical sophistication between the six writing prompts.

Figure 5.6: Means plots of the differences in lexical sophistication between the six writing prompts



The figure clearly shows that Prompt 115 has a much lower mean factor score for lexical sophistication than the other prompts. The figure also indicates that Prompts 100, 108, and 214 have higher means for lexical sophistication than for the other prompts. The differences between Prompt 115 and Prompts 100, 108, and 214 are all significant at the 95% level.

Table 5.14 shows the findings of the post-hoc test (Bonferroni) for factor score 4 (academic vocabulary use). The data show that there are significant differences in academic vocabulary use between the following prompts for factor score 1 ($p<.5$).

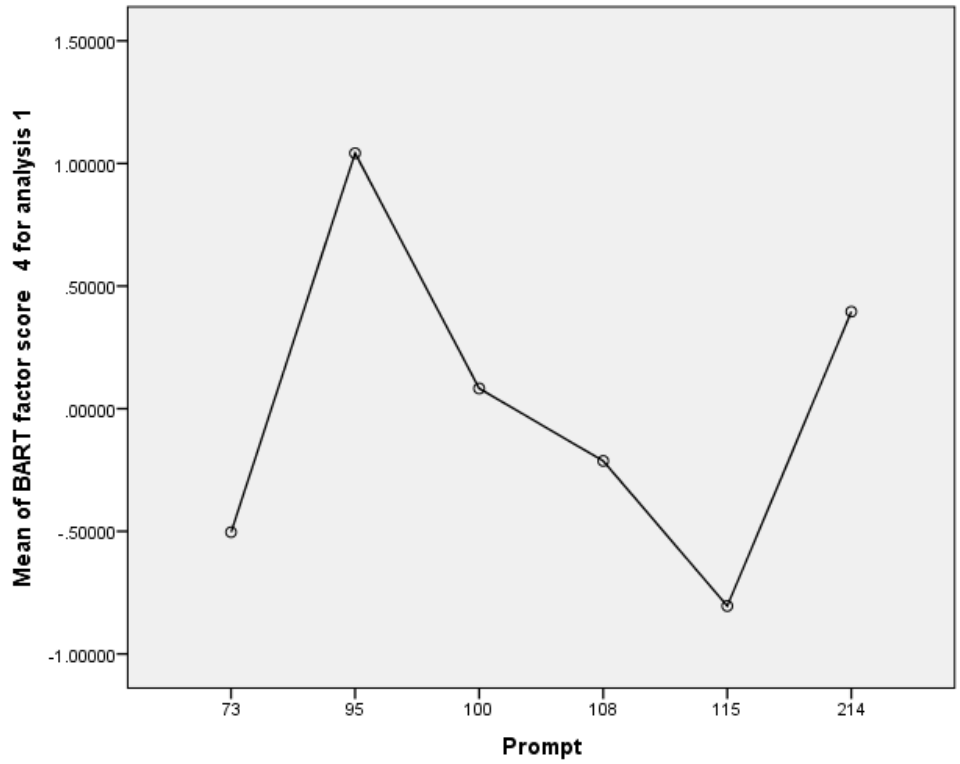
- 73 (mistakes) < 95 (child psychologists)
- 73 (mistakes) < 100 (professions)
- 73 (mistakes) < 214 (government leaders)
- 95 (child psychologists) > all other prompts
- 100 (professions) > 115 (memorable days)
- 108 (standard work week) > 115 (memorable days)
- 108 (standard work week) < 214 (government leaders)
- 115 (memorable days) < 214 (government leaders)

Table 5.14: Post-hoc test for factor score 4 (academic vocabulary use)

Prompt			Mean Difference	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Bonferroni	73	95	-1.545*	0.15	.00	-1.98	-1.11
		100	-.586*	0.15	.00	-1.02	-.15
		108	-.290	0.15	.72	-.72	.14
		115	.300	0.15	.61	-.13	.73
		214	-.899*	0.15	.00	-1.33	-.47
	95	73	1.545*	0.15	.00	1.11	1.98
		100	.959*	0.15	.00	.53	1.39
		108	1.255*	0.15	.00	.82	1.69
		115	1.846*	0.15	.00	1.41	2.28
		214	.646*	0.15	.00	.21	1.08
	100	73	.586*	0.15	.00	.15	1.02
		95	-.959*	0.15	.00	-1.39	-.53
		108	.300	0.15	.66	-.14	.72
		115	.887*	0.15	.00	.45	1.32
		214	-.310	0.15	.50	-.75	.12
	108	73	.290	0.15	.72	-.14	.72
		95	-1.255*	0.15	.00	-1.69	-.82
		100	-.300	0.15	.66	-.73	.14
		115	.591*	0.15	.00	.16	1.02
		214	-.609*	0.15	.00	-1.04	-.18
	115	73	-.300	0.15	.61	-.73	.13
		95	-1.846*	0.15	.00	-2.28	-1.41
		100	-.887*	0.15	.00	-1.32	-.45
		108	-.591*	0.15	.00	-1.02	-.16
		214	-1.200*	0.15	.00	-1.63	-.77
	214	73	.899*	0.15	.00	.47	1.33
		95	-.646*	0.15	.00	-1.08	-.21
		100	.310	0.15	.50	-.12	.75
		108	.608*	0.15	.00	.18	1.04
		115	1.200*	0.15	.00	.77	1.63
*. The mean difference is significant at the .05 level.							

Figure 5.7 shows the means plots of the differences in factors score 4 (academic vocabulary use) between the six writing prompts.

Figure 5.7: Means plots of the differences in academic vocabulary use between the six writing prompts



The figure shows that Prompt 95 has the highest mean value for academic vocabulary use and the mean for Prompt 95 is significantly different from the mean for all other prompts. The figure also reveals that Prompts 73 and 115 have the lowest means for academic vocabulary use.

Table 5.15 shows the findings of the post-hoc test (Bonferroni) for factor score 5 (syntactic complexity). The data show that there are significant differences in syntactic complexity between the following prompts for factor score 5 ($p<.5$).

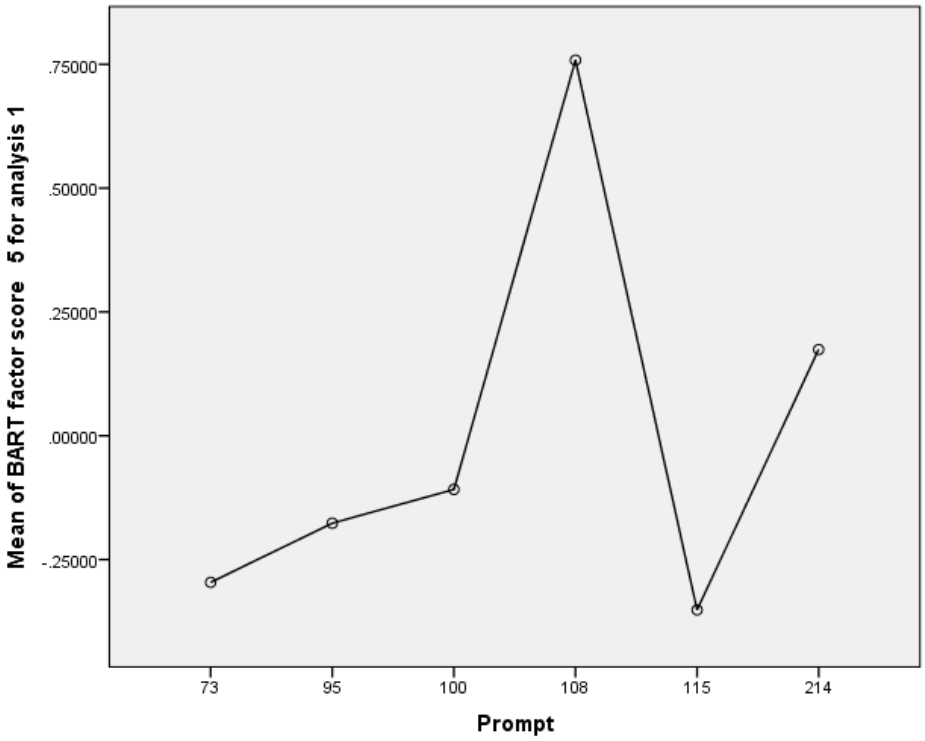
- 108 (standard work week) > all other prompts
- 115 (memorable days) < 214 (government leaders)

Table 5.15: Post-hoc test for factor score 5 (syntactic complexity)

Prompt			Mean Difference	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Bonferroni	73	95	-.120	.17	1.00	-.62	.38
		100	-.190	.17	1.00	-.69	.32
		108	-1.054*	.17	.00	-1.56	-.55
		115	.060	.17	1.00	-.45	.56
		214	.470	.17	.09	-.97	.03
	95	73	.120	.17	1.00	-.38	.62
		100	-.070	.17	1.00	-.57	.43
		108	-.935*	.17	.00	-1.44	-.43
		115	.170	.17	1.00	-.33	.68
		214	-.350	.17	.60	-.85	.15
	100	73	.190	.17	1.00	-.32	.69
		95	.070	.17	1.00	-.43	.57
		108	-.867*	.17	.00	-1.37	-.36
		115	.240	.17	1.00	-.26	.75
		214	-.280	.17	1.00	-.79	.22
	108	73	1.054*	.17	.00	.55	1.56
		95	.935*	.17	.00	.43	1.44
		100	.867*	.17	.00	.36	1.37
		115	1.110*	.17	.00	.61	1.61
		214	.584*	.17	.01	.08	1.09
	115	73	-.060	.17	1.00	-.56	.45
		95	-.170	.17	1.00	-.68	.33
		100	-.240	.17	1.00	-.75	.26
		108	-1.110*	.17	.00	-1.61	-.61
		214	-.525*	.17	.03	-1.03	-.02
	214	73	.470	.17	.09	-.03	.97
		95	.350	.17	.60	-.15	.85
		100	.280	.17	1.00	-.22	.79
		108	-.584*	.17	.01	-1.09	-.08
		115	.528*	.17	.03	.02	1.03
*. The mean difference is significant at the .05 level.							

Figure 5.8 shows the means plots of the differences in factors score 5 (syntactic complexity) between the six writing prompts.

Figure 5.8: Means plots of the differences in syntactic complexity between the six writing prompts



The figure shows that Prompt 108 has the highest mean for factor score 5 (syntactic complexity) and that the mean for Prompt 108 is significantly different from the other mean factors scores. Prompt 115 has the lowest mean factor score for syntctic complexity but this is signicantly different from only Prompts 108 and 214.

The findings from the post-hoc tests show that several of the writing prompts conribute to the significant differences revealed by the MANOVA and ANOVA analyses. A discussion of the prompts that contribute to the significant differences identifed in the factor scores will be presented in Section 5.8.

5.7 Summary of results

The results of the MANOVA and ANOVA analyses indicate that there are significant effects of writing prompt features on:

- Lexical sophistication
- Cohesion
- Academic vocabulary use
- Syntactic complexity
- Fluency

Of these traits of second language writing proficiency, the largest effect sizes came from lexical sophistication, academic vocabulary use, and syntactic complexity. Several of the writing prompts were seen to contribute to these observed effects. There was no significant effect of prompt features on the accuracy of the written language or on the writing score.

Overall, these findings present a broader range of significant effects as a result of prompt characteristics than has been seen in previous research from the second language writing literature. A full discussion of the findings reported above and how they compare to the findings of comparable studies in the literature is presented in Chapter 8. The reasons for the relatively large number of significant effects in the current work may be due to several issues but the systematic differences in prompt characteristics (response mode, domain, number of rhetorical cues, and the focus of the prompt), relatively large sample size, and selection of certain discourse measures have probably contributed to the findings reported above. Section 5.8 considers, in detail which of the writing prompts most contribute to the observed significant effects in an attempt to identify the distinguishing prompt characteristics that may contribute to prompt effect.

5.8 Relationships between prompt characteristics and significant differences in written products

Numerous significant differences were identified in the discourse measures used to analyze the essays written in response to the six different writing prompts: writing prompt characteristics have a significant effect on the written language elicited. This finding indicates that the writing prompts analyzed in this study are not equivalent. The prompts elicit written products that differ significantly in terms of several textual features.

The purpose of this section is to explore some possible reasons for the differences. This will focus on examining which prompt characteristics are responsible for eliciting responses with significantly different textual features. The MANOVA and post-hoc analyses revealed that there were significant effects of the writing prompt characteristics on the lexical sophistication, academic vocabulary use, syntactic complexity, cohesion, and fluency of the responses to the prompts. The largest effect sizes were shown in the discourse measures that operationalized the textual features of lexical sophistication, academic vocabulary use, and syntactic complexity.

The results of the factor analysis and MANOVA post-hoc tests show that three factors (lexical sophistication, academic vocabulary use, and syntactic complexity) accounted for the largest proportion of total variance. Hence, the investigation of the prompts and characteristics that contribute to prompt effect will initially focus on these three factors.

Before beginning the detailed discussion of prompts that contribute to the observed significant differences in discourse measures, the six prompts and how they were categorized are reproduced below in Table 5.16. This table is a reproduction of Table 4.4.

Table 5.16: Prompts used in main study

Prompt #	**Prompt ID**	**Keywords**	**Prompt**
1	95	child psychologists	Some child psychologists believe that the peer groups children play with influence their character and personality development more than the children's parents do. The psychologists say children are more interested in fitting in with their friends than behaving the way their parents want them to. Do you agree or disagree with these psychologists? Explain your point of view.
2	214	government leaders	How important is it to know about the personal life (e.g. health, personal relationships, youthful mistakes) of government leaders? What things should be made public? What things should be kept private? Give your opinion and support it with reasons.
3	100	professions	Would you rather have the same profession all your life or change jobs often? Explain the reasons for your preference.
4	108	standard work week	In some countries such as the United States, the standard work week is 5 days, 8 hours a day. Some companies are considering changing their work week to 4 days, 10 hours a day. What are the advantages and disadvantages of the two options? Which do you think would be better for society? Give reasons to support your opinion.
5	115	memorable days	In everyone's life, some days are particularly memorable. Describe what happened on one such day that you have experienced in your own life. What made it memorable and why? Include specific details such as where you were, who else was there, and so on, that will convince the reader how memorable the day was.
6	73	Mistakes	It is often said that we learn more from our mistakes than our successes. Tell about a mistake that you once made and learned something from.

5.8.1 Lexical sophistication

The MANOVA and discriminant function analyses revealed significant differences in the factor scores for the discourse measures that operationalize lexical sophistication. The MANOVA post-hoc tests showed

that there were significant differences in the factor scores for lexical sophistication in the responses elicited by the following prompts:

- 73 (mistakes) > 100 (professions)
- 73 (mistakes) > 108 (standard work week)
- 73(mistakes) > 214 (government leaders)
- 95 (child psychologists) < 115 (memorable days)
- 100 (professions) < 115 (memorable days)
- 108 (standard work week) < 115 (memorable days)
- 115 (memorable days) > 214 (government leaders)

Prompt 115 (memorable days) elicited responses that scored lowest on lexical sophistication. Prompt 73 (mistakes) also elicited responses with significantly lower factor scores for lexical sophistication than responses to some other prompts (100, 108, and 214). Prompts 100, 108, and 214 all elicited responses that had factors scores that were very similar to each other but significantly higher than those of responses to other prompts.

The following section explores differences in descriptive statistics for each of the variables that operationalize lexical sophistication to see how the prompts listed above differ.

Table 5.17 shows the descriptive statistics for the lexical frequency profile by prompt

Table 5.17: Descriptive statistics for lexical frequency profile

FREQ1	**Prompt**	**Mean**	**Std. deviation**	**N**
	73	79.12	4.72	60
	95	80.17	4.31	60
	100	81.55	4.59	60
	108	81.47	3.44	60
	115	76.83	3.58	60
	214	81.92	4.09	60
	Total	80.18	4.49	360
FREQ2	**Prompt**	**Mean**	**Std. deviation**	**N**
	73	11.9	2.7	60
	95	10.78	2.66	60
	100	9.55	3.07	60
	108	10.6	2.76	60
	115	12.35	2.5	60
	214	10.5	2.61	60
	Total	10.94	2.86	360
FREQ3	**Prompt**	**Mean**	**Std. deviation**	**N**
	73	9.07	2.99	60
	95	9.08	2.4	60
	100	8.97	2.72	60
	108	7.97	2.13	60
	115	10.7	2.36	60
	214	7.55	2.65	60
	Total	8.89	2.73	360

The FREQ1 data shows the percentage of words in a response from the 500 most frequent words of English in the COCA corpus. Hence, responses with a high percentage of FREQ1 are not lexically sophisticated. The FREQ2 data shows the percentage of a response that is made up of the 501 to 3,000 most common words within the COCA corpus. The FREQ3 data shows the percentage of a response that is made up of words beyond the 3,000 most common words within the COCA corpus. Hence, responses with a high percentage of FREQ3 are lexically sophisticated.

The descriptive statistics for FREQ1 show that prompt 115 (memorable day) elicited the lowest percentage of high-frequency vocabulary (76.8%). The other prompts elicit between 79.1% and 81.9% of high-frequency vocabulary, with prompt 214 (government leaders) eliciting the highest percentage of high-frequency vocabulary. As would be expected, these results are reversed for the FREQ3 data. Prompt 115 elicits the highest percentage of low-frequency vocabulary (10.7%) and prompt 214 elicits the lowest percentage of low-frequency vocabulary (7.55%). Prompt 108 (standard work week) also elicits a noticeably lower percentage (7.97%) than the other prompts.

The FREQ2 data is less straightforward to interpret as this frequency band contains neither highly frequent words as in FREQ 1 nor infrequent words as in FREQ3. Rather, it includes the core vocabulary that language learners and developing writers need to acquire as they progress from a beginner level of second language development. However, the FREQ2 data shows quite clearly that prompts 73 and 115 elicit higher percentages of FREQ2 vocabulary than do the other prompts. Prompt 100 (professions) elicits the lowest percentage of FREQ2 vocabulary. The excerpts from essays described in section 6.1.1 will aid with the interpretation of FREQ2 data.

Table 5.18 shows the descriptive statistics for type-token ratio by prompt.

Table 5.18: Descriptive statistics for type-token ratio

Prompt	Mean	Std. deviation	N
73	0.55	0.06	60
95	0.54	0.05	60
100	0.53	0.09	60
108	0.51	0.08	60
115	0.54	0.05	60
214	0.54	0.06	60
Total	0.53	0.07	360

The post-hoc tests for the factor score that operationalized lexical sophistication indicated that there was a significant effect of prompt characteristics on the lexical sophistication of the responses. However, that is not reflected in the type-token ratio data. There are relatively few differences between the TTR means for the six prompts. The only prompt that differs significantly from any of the other prompts is prompt 108 (standard work week). It has a TTR that is significantly lower than the TTR of prompt 73 (mistakes). These results indicate that Prompt 108 elicits language that is on average less lexically sophisticated than in responses to other prompts. However, the differences between the means of the other prompts are not significant, indicating that TTR is contributing relatively little to the significant differences revealed in lexical sophistication between the responses.

The lexical frequency profile data showed that prompts 115 and 73 elicit vocabulary that is more lexically sophisticated than the other prompts. Prompts 214 (government leaders) and 108 (standard work week) elicit vocabulary that is less lexically sophisticated than the other prompts. The type-token ratio data also indicated that prompt 108 elicits language that is less lexically sophisticated than the responses to other prompts. In summary these prompts do not elicit language that is similar or comparable in terms of lexical sophistication. A discussion of the prompt characteristics that most contribute to the differences in lexical sophistication will follow later in this chapter in Section 5.8.4.

5.8.2 Academic vocabulary use

The ANOVA and discriminant function analyses revealed significant differences in the factors scores for the discourse measures that operationalize academic vocabulary use. The MANOVA post-hoc tests

showed that there were significant differences in the factor scores for academic vocabulary use in the responses elicited by the following prompts:

- 73 (mistakes) < 95 (child psychologists)
- 73 (mistakes) < 100 (professions)
- 73 (mistakes) < 214 (government leaders)
- 95 (child psychologists) > all other prompts
- 100 (professions) > 115 (memorable days)
- 108 (standard work week) > 115 (memorable days)
- 108 (standard work week) < 214 (government leaders)
- 115 (memorable days) < 214 (government leaders)

Prompts 73 (mistakes) and 115 (memorable day) elicited responses with the lowest factor scores for academic vocabulary use. Prompt 95 (child psychologists) elicited responses with a much higher factor score for academic vocabulary use than the responses to other prompts. The discourse measures that operationalized academic vocabulary use were revealed through the factor analysis to be the vocabulary classified as academic within the COCA corpus along with average word length. The following section explores differences in descriptive statistics for each of the two variables that operationalize academic vocabulary use to see how the prompts listed above differ.

Table 5.19 shows the descriptive statistics for the percentages of words classified as academic within the COCA corpus.

Table 5.19: Percentages of words classified as academic within the COCA corpus

Descriptive Statistics

	Prompt	Mean	Std. Deviation	N
FREQ AC	73	4.20	1.981	60
	95	8.25	3.887	60
	100	6.23	3.164	60
	108	5.18	2.332	60
	115	2.47	1.741	60
	214	5.27	2.442	60
	Total	5.27	3.208	360

Prompt 115 (memorable days) elicited a noticeably low percentage of academic vocabulary (2.47%). Prompt 73 (mistakes) also elicited relatively low percentages of academic vocabulary (4.2%). Prompt 95 (child psychologists) stands out as eliciting the highest percentage (8.25%) of academic vocabulary of all the prompts. Prompt 100 (professions) also elicited a relatively high percentage of academic vocabulary.

Table 5.20 shows the descriptive statistics for average word length.

Table 5.20: Descriptive statistics for average word length

Descriptive Statistics

	Prompt	Mean	Std. Deviation	N
AWL	73	4.092	0.198	60
	95	4.588	0.223	60
	100	4.240	0.273	60
	108	4.156	0.252	60
	115	4.093	0.200	60
	214	4.485	0.287	60
	Total	4.276	0.308	360

Prompts 95 (child psychologists) elicited responses with the longest average word length (4.588 letters). Prompt 214 (government leaders) also elicited responses with noticeably longer average word length (4.485 letters) than other prompts. Prompts 73 (mistakes) and 115 (memorable days) elicited responses with the shortest average word lengths (4.092 and 4.093 letters, respectively).

The data for the two discourse measures that operationalize academic vocabulary use show some consistent findings. Prompts 73 (mistakes) and 115 (memorable days) both elicited low percentages of words that are classified as academic within the COCA corpus and elicited responses with the shortest average word length. Interestingly though, these two prompts also elicited responses that are lexically more sophisticated than the other prompts. While these prompts elicited relatively high percentages of low-frequency vocabulary, they also elicited low percentages of academic vocabulary. Prompts 73 and 115 also elicited responses that have the shortest average word length among the six prompts. Hence, in terms of lexical sophistication and academic vocabulary use, these two prompts (73 and 115) stand out as eliciting responses that are markedly different from the other four prompts in the study.

Prompt 95 (child psychologists) elicited responses that have both the highest percentage of academic vocabulary and the longest average word length. This is the one prompt that stands out from the others in consistently eliciting more academic language use than the other prompts. Prompt 95 also elicited a relatively low percentage of high-frequency vocabulary (80.167% FREQ1) and a relatively high percentage of low-frequency vocabulary (9.083% of FREQ3). However, it is the academic vocabulary use that stands out for prompt 95. The data indicate that these prompts (particularly 73, 115, and 95) did not elicit language that is similar or comparable in terms of academic vocabulary use. The reasons for these results are discussed and accounted for in Chapter 6.

5.8.3 Syntactic complexity

The ANOVA and discriminant function analyses revealed significant differences in the factors scores for the discourse measures that operationalize syntactic complexity. The two variables that operationalized syntactic complexity, as identified through the factor analysis were syntactic left embeddedness (the average number of words before the first verb in a sentence) and average sentence length. The

MANOVA post-hoc tests showed that there were significant differences in the factor scores for syntactic complexity in the responses elicited by the following prompts:

- 108 (standard work week) > all other prompts
- 115 (memorable days) < 214 (government leaders)

Prompt 108 (standard work week) had the highest factor score for syntactic complexity, with the prompt eliciting language that was significantly more syntactically complex than the language elicited by all other prompts. Prompts 73 (mistakes) and 115 (memorable days) elicited responses with the lowest factor scores for syntactic complexity. The other prompts elicited language that was similar in terms of syntactic complexity. The following section will explore the descriptive statistics for these two variables to examine how specific prompts differed by these two variables.

Table 5.21 shows the descriptive statistics for syntactic left embeddedness

Table 5.21: Descriptive statistics for syntactic left embeddedness

Descriptive Statistics

Prompt		Mean	Std. Deviation	N
SYNLE	73	3.738	1.523	60
	95	4.058	1.448	60
	100	4.327	2.006	60
	108	5.450	1.453	60
	115	3.831	1.126	60
	214	4.148	1.335	60
	Total	4.259	1.599	360

As indicated by the MANOVA post-hoc tests, prompt 108 (standard work week) elicited language with significantly more words before the first verb (5.45 words) than the other prompts, an indicator of syntactically complex sentences. Prompts 73 (mistakes) and 115 (memorable days) elicited responses with fewer words (3.74 words and 3.83 words respectively), on average before the first verb than the other prompts.

Table 5.22 shows the descriptive statistics for average sentence length.

Table 5.22: Descriptive statistics for average sentence length

Descriptive Statistics

Prompt		Mean	Std. Deviation	N
ASL	73	16.578	4.264	60
	95	17.094	4.334	60
	100	17.089	5.185	60
	108	20.911	4.757	60
	115	16.166	4.139	60
	214	19.529	4.680	60
	Total	17.895	4.858	360

Prompt 108 (standard work week) elicited responses with longer sentences, on average than other prompts. Prompt 214 (government leaders) also elicited responses with noticeably longer sentences than the other prompts. Prompts 115 (memorable days) and 73 (mistakes) elicited responses with shorter average sentences than other prompts.

The data indicate that these six prompts do not elicit language that is similar or comparable in terms of syntactic complexity. One prompt (108) elicits language that is significantly more complex than the language produced in response to other prompts, as shown by both discourse measures that operationalize syntactic complexity. Prompts 73 (mistakes) and 115 (memorable days) elicit language that is syntactically less complex than the other prompts, again based on both indicators of syntactic complexity. The reasons for these differences are discussed in Chapter 6.

5.8.4 Summary of prompts that elicit different written products

The picture painted by the results of the quantitative analyses of the discourse measures is that there are numerous significant differences in the responses elicited by the six writing prompts. Table 5.23 shows a summary of how the responses elicited by the six writing prompts differ.

Table 5.23: Summary of significant differences elicited by prompts

Prompt #	Prompt key words	Significant differences
95	child psychologists	High academic vocabulary use
214	government leaders	Low lexical sophistication High academic vocabulary use High syntactic complexity
100	professions	None
108	standard work week	Low lexical sophistication High syntactic complexity
115	memorable days	High lexical sophistication Low academic vocabulary use Low syntactic complexity
73	Mistakes	High lexical sophistication Low academic vocabulary use Low syntactic complexity

The summary shows that while some prompts elicit responses with high levels of lexical sophistication (115 and 73), the same prompts elicit language with low levels of academic language use and syntactic complexity. Conversely, other prompts (214 and 108) elicit responses with low levels of lexical sophistication but with high levels of syntactic complexity and in the case of prompt 214 high levels of academic vocabulary use. These two pairs of prompts present a valuable case study to help understand which prompt characteristics contribute to the observed differences in the written product. Table 5.24 summarizes the characteristics of the four prompts that elicit the significant differences in discourse measures described above.

Table 5.24: Characteristics of prompts that elicit responses with significant differences in written product

Prompt #	Domain	Response mode	# of rhetorical cues	Focus
214	Public	Argumentative	5	Focused
108	Occupational	Argumentative	6	Focused
115	Personal	Narrative	6	Focused
73	Personal	Narrative	2	Open

It is immediately apparent that prompts 73 and 115, (high levels of lexical sophistication, low levels of academic vocabulary use and syntactic complexity) are both situated in the personal domain and elicit narrative responses. These are the distinguishing characteristics of these two prompts. Prompts 214 and 108 (low lexical sophistication, high syntactic complexity) also share some characteristics; they are both

focused prompts, with a relatively large number of rhetorical and cues that elicit argumentative responses.

The two prompt characteristics that clearly differentiate the two pairs of prompts (73 and 115 from 214 and 108) are domain and response mode. Prompts 73 and 115 are both situated in the personal domain and neither prompt 214 nor prompt 108 are situated in the personal domain. Prompts 73 and 115 both elicit narrative responses, while prompts 214 and 108 both elicit argumentative responses. Hence, domain and response mode appear to be prompt characteristics that contribute to identifiable differences in written product. The extent to which the other two prompt characteristics (number of rhetorical cues and focus) contribute to the differences in written product is less clear. Three of the four prompts identified above have relatively large numbers of rhetorical cues and are focused. This may be indicative that these characteristics contribute to prompt effect but the nesting within the research design presents a challenge to interpreting the importance of these characteristics in comparison with domain and response mode.

As reported in Section 4.3, these six writing prompts were administered on live test forms and the responses analyzed were written by authentic test takers. While the nesting of prompt characteristics within the six prompts precludes definitive conclusions regarding which prompt characteristics may be responsible for the significant differences in the written product, the fact that these prompts and responses are from live test administrations and that the test takers would have been fully motivated to demonstrate the full range of their language resources on the test, means that the findings of this study are meaningful and potentially generalizable to other second language writing tests that employ a variety of independent writing tasks.

The results of the MANOVA and post-hoc tests provided evidence that the written responses to different prompts differ significantly in relation to key textual features. In the next chapter, excerpts from the responses to the prompts will be presented to exemplify how the responses differ. The aim of providing the excerpts is to contextualize how differences in lexical sophistication, academic vocabulary use, and syntactic complexity manifest themselves in authentic test taker writing. The excerpts from different prompts will allow for a better understanding of how those differences are manifested in the actual writing of test takers.

Chapter 6 – Discussion of findings from main study

The results reported in the previous chapter showed that the six independent writing prompts studied elicit written language that differs significantly in terms of a range of textual features. The textual features that varied, based on the prompt responded to, were:

- Lexical sophistication
- Cohesion
- Academic vocabulary use
- Syntactic complexity
- Fluency

Of these textual features, the largest effect sizes were observed for lexical sophistication, academic vocabulary use, and syntactic complexity. However, there were no significant differences observed in accuracy or in the holistic score awarded to the responses.

In this chapter, these findings will be discussed. In addition, samples of the written responses will be provided to help exemplify the significant differences in the responses. These sample responses will provide some contextualization for the findings by demonstrating what the significant statistical differences look like in authentic test taker writing.

6.1 Summary of quantitative results

The findings presented and discussed in this chapter will directly address the second research question in this work.

> How do these [prompt] characteristics affect the test-takers' final written product?

The findings will be discussed in the context of the relationship between the prompt characteristics (domain, response mode, focus, number of rhetorical cues) and the textual features of the written responses, the relationships that are central to the quantitative study in this thesis.

The results reported in Chapter 5 showed that there are some clear relationships between prompt characteristics and the textual features of the responses (see Table 5.24 on p.113). The results indicated that the prompt characteristics of domain and response mode most clearly elicit responses with significant differences in textual features. There is also evidence that the prompt characteristics of focus and number of rhetorical cues elicit responses with significant differences in textual features, but the results are less straightforward to interpret.

6.1.1 Domain and response mode

Two prompt characteristics that are clearly associated with significantly different written products are domain and response mode (see Chapter 5, Section 5.8.4):

- Responses to prompts situated in the personal domain that elicit a narrative response are characterized by high levels of lexical sophistication, low levels of academic language use, and low levels of syntactic complexity.
- Responses to prompts situated in non-personal domains and that elicit an argumentative response are characterized by low levels of lexical sophistication and high levels of syntactic complexity.

6.1.2 Focus and number of rhetorical cues

The effect of these prompt characteristics on the written products is less clear than for the characteristics of domain and response mode (see Section 5.8.4). The results of the quantitative analyses by themselves do not allow for clear conclusions to be drawn about the relationships between these prompt characteristics and the textual features of the written responses. A majority of the prompts that most contribute to responses with significant differences in written products are focused and have a large number of rhetorical cues. However, the textual features of responses that are associated with these prompt characteristics are not yet clear. The quantitative analyses in Chapter 5 cannot provide a complete understanding of the relationships between the prompt characteristics of focus and number of rhetorical cues and the written products elicited from these characteristics. It will be necessary to look more closely at the responses in order to better understand the effect of these prompt characteristics.

The following sections will present excerpts from the responses to the prompts studied in this work. The aim of presenting these responses is to help clarify how the significant differences in written products are manifested in authentic samples of test taker performances. These samples help move beyond a reliance on purely quantitative analyses and help provide a more contextualized understanding of how the responses to certain prompts differ from responses to other prompts.

6.2 Contextualization of the significant differences in written products

This section will provide excerpts from authentic responses that exemplify the differences in the written product by textual feature and proficiency level. This will allow for a further insight into how prompt effect can be manifested in responses to prompts that differ by particular prompt characteristics.

In order to look deeper into the specific textual differences in the written products elicited by particular prompts, the following sections will describe trends in the essays, via full responses and excerpts from individual writers within the dataset. It is hoped that returning to the texts themselves will shed additional light on how the written products vary.

6.3 Excerpts from responses to prompts

The excerpts from the written products will be presented by prompt and by proficiency level. The aim of this work is to identify specific prompt characteristics that elicit responses that differ significantly from responses to prompts with different characteristics. Excerpts from specific prompts (based on the significant differences reported in Chapter 5 and summarized in Tables 5.24 and 5.25) will be presented from writers at high, medium, and low levels of overall language proficiency to help exemplify how the written products vary, within a defined ability range by prompt.

In the writing samples, an entire response is first presented and this is then followed by shorter excerpts. The full response provides an understanding of the writing proficiency of the test population at the high, medium, and low levels of proficiency within the sample population. The shorter excerpts provide examples of particular textual features that are relevant for the responses to specific prompts. For example, responses to prompts 73 and 115 are characterized by high levels of lexical sophistication, so the excerpts provided in this section exemplify this textual feature.

Before each excerpt is presented, relevant discourse measure data is shown that illustrates the textual features of each response. This data shows the textual features that most clearly represent the textual features of lexical sophistication, academic vocabulary use, and syntactic complexity:

- The percentage of low-frequency (lexically sophisticated) vocabulary in the response (FREQ3)
- The percentage of academic vocabulary in the response (FREQAC)
- The average number of words before the first verb in a sentence (SYNLE)
- The average sentence length (ASL)
- Writing score awarded to the response

The justifications for these discourse measures and details of how they were calculated was provided in Chapter 4 (p.73-75).

6.3.1 Responses to Prompts 73 and 115

> It is often said that we learn more from our mistakes than our successes. Tell about a mistake that you once made and learned something from. (Prompt 73)
>
> In everyone's life, some days are particularly memorable. Describe what happened on one such day that you have experienced in your own life. What made it memorable and why? Include specific details such as where you were, who else was there, and so on, that will convince the reader how memorable the day was. (Prompt 115)

These prompts are both situated in the personal domain and elicit narrative responses. The responses to both these prompts are characterized by high levels of lexical sophistication but low levels of academic language use and syntactic complexity. The low-frequency vocabulary (FREQ3) within these responses has been bolded and underlined.

6.3.1.1 Responses from the high-proficiency group

The following excerpts are drawn from the high-proficiency band of the test takers who responded to these prompts (73 and 115).

Excerpt 1_49144115 (Prompt 115, Writing score 83)

(L1 Kashmiri, FREQ3 15%, FREQAC 3%, SYNLE 2.64, ASL 12.56)

> *The most **memorable** days of my life are the days when I got married. It was in January 2000. Those days are really saved and **treasured** in my mind too. Those days gave me very big happiness and of course some sadness too.*
>
> *Me and my husband decided to get married in January 2000. We both were working in Saudi Arabia. So we had limited days of **vacation**. We both had to visit our parents and do the necessary **preparations** for our marriage.*
>
> *2nd thing was that we both are from 2 different **nationalities**. I am Indian and my husband is Pakistani. He went to Pakistan from Saudi Arabia and applied for a **visa** for India in Pakistan only. That was because he had to apply for his parents **visa** also. So his parents and him got the visa for India only for 7 days. They had to go to India before January 20 so that they can go back to Pakistan before January 26. which is Republic day of India that was because of that **prevailing** circumstances between 2 countries. They got **visa** only to Delhi. They were not allowed to go **anywhere** else in India. I am from Northern part of India called Kashmir (The distonted **territory** between the two countries). So we had to travel from Kashmir to Delhi. Me my parents and my relatives prepared for **journey** to Delhi by bus. Every thing was ready, our lugages and every necessary things was loaded on the bus. We decided to leave early morning on January 18. So on January 17, 2000 everybody slept ready to wakeup in the morning to leave for Delhi.*
>
> *On January 18, 2000 everybody woke up only to see whole the valley of Kashmir covered in 3 feet, white **blanket** of snow. All the roads were blocked. The valley was cut off from the rest of the country. Flights were **cancelled**. My would be husband and in laws were already in Delhi waiting for us. Flights were being **rescheduled**. It was hard to get **plain** tickets. My **uncles** somehow **arranged** 2 tickets for me and my father. So both of us flew to Delhi. Rest of my family was in Kashmir. My parents and **siblings** were happy that I was getting married, but they were sad also that they were not there with me. So we got married on Jan 20, 2000. I was happy to find the partner that I had really wished to be with but also was **sad** that nobody other than my father was able to participate in my marriage.*
>
> *So these are the **bitter** sweet memories of my life which I am going to **treasure** for rest of my life.*

Excerpt 2_4884773 (Prompt 73, Writing score 75)

(L1 Mandarin, FREQ3 13%, FREQ AC 6%, SYNLE 3.16, ASL 11.4)

> Before the final **contest***, those who failed to go into the final* ***contest*** *would be* ***rewarded*** *by 1000 Yuan and* ***testimonials****.* ***Via*** *the email I* ***informed*** *them to* ***fetch*** *the* ***rewards****. How* ***careless*** *I was. I sent the information to an extra group who was not* ***rewarded****. Suddenly, my brain was* ***blank****. I told it to my leader. He was a bit angry and shouted at me "How comes? What a* ***careless*** *guy you are." I nodded my head and showed my responsibility for the whole thing. Due to the good name of the organization I paid for extra 1000 Yuan to* ***reward*** *the team. This thing* ***impress*** *me in my whole life.*

Excerpt 3_49161115 (Prompt 115, Writing score 83)

(L1 Spanish, FREQ3 13%, FREQAC 3%, SYNLE 3.23, ASL 13.83)

> *It has been almost three years when my* ***eldest*** *daughter had her* ***scolioris*** *surgery. It was a very* ***scary*** *moment for our family since she was only 17 years old when she had the procedure.* ***Prior*** *to the surgery, she was* ***diagnosed*** *that she has a* ***moderate*** ***scolioris****, this was when we were still back home. When we landed here in Canada, we took her to a* ***specialist*** *for further* ***follow-up****.*

The personal content in the responses to these two prompts is apparent from these excerpts. They address several different topics and low-frequency vocabulary is commonly used (*prevailing, territory, testimonials, impress, erased, blanket of snow, diagnosed, rescheduled*). In many cases, the low-frequency vocabulary is used appropriately within the context of the response.

The open focus of the prompts allows test takers to write on a very broad range of topics and still respond on topic. Additionally, as the prompts are situated in the personal domain, the test taker is invited to draw on personal experience. Both these prompt characteristics (open focus and personal domain) contribute to the lexically sophisticated responses that were elicited as the writers were describing familiar experiences. The ability to select a topic that is personally familiar to the writer means that the individual is able to draw on lexical resources that he or she has likely used previously to recount a particular event. The combination of the open focus and the personal domain may advantage the test taker over those who respond to other prompts as the test taker is able to marshal the full range of his or her lexical resources.

Although these responses are more lexically sophisticated than responses to other prompts, even the high-proficiency writers tend not to write on academic topics or utilize a high percentage of academic vocabulary. Excerpt 3, above is an exception as the writer included some field-specific, academic terms in the response. For the most part though, even though these writers probably have a broad vocabulary to draw from, they produce relatively little academic vocabulary in response to prompts 73 and 115.

6.3.1.2 Responses from the medium proficiency group

The following excerpts are drawn from the medium-proficiency band of the test takers who responded to prompts 73 and 115.

Excerpt 4_49180115 (Prompt 115, Writing score 73)

(L1 Punjabi, FREQ3 13%, FREQAC 2%, SYNLE 5.86, ASL 17.67)

*Everyperson has few **memorable** days in their life that make them to feel good, to laugh, to **relax**. That day may be **memorable** due to some good news or some bad but in most of the people's life the good things goes in the memories. Same thing is with me. Mine **memorable** day is 4 oct, 2001, the **birthday** of my nephew. **Whenever** I think about that day my eyes glass with **joy***

*I was in India at my own birth place with my whole family. I have two sisters but no brother. Like in every community, it is very important to have a son in the family who will lead the family, same thing is there in my community also. People always talk about my family that who will going to be the leader of the family. But on the day of birth of my **nephew** everyone in my family seemed to be **relaxed** as after 40 years a boy took birth in my family.*

*When we got the news we started **distributing** sweets to our relatives like there was a **festival**. All my family members and relatives gathered at my place to share this **joy**. The head **priest** from my religious **temple** came to gave his **blessing** to my **nephew**.*

*Moreover, we did the **fireworks** at the night same as we done on the day of 'Diwali' can Indian **festival**; we did dancing, **fireworks** and enjoyed a lot on that day. My parents and **grandparents** were extremely happy. Everybody could see the sign of **joy** and a kind of relief on their faces. That day became more remembrable when we heard another good news that my dad got a **promotion** in his job. All my relatives said that my **nephew** is very **lucky** for the family. They said that his birth will bring **happiness** and **glory** to the family and **fortunately** this thing happens.*

***Lastly** I want to say that the **birthday** of my **nephew** is **unforgetable** day for me, may be because of the blood is thicker than the water. **Whenever** I open my photo **album** for that day or just remember that day all activities that happened just plays like a movie in front of me.*

Excerpt 5_5076573 (Prompt 73, Writing score 73)

(L1 Tagalog, FREQ3 17%, FREQAC 5%, SYNLE 3.17, ASL 13.67)

*I **resigned** my job, was working in University. The work was light, no **tension**. Actually I had been earning **salary** without working when planning to **migrate** to Canada, work experience is a necessary requirement. This is the reason of **resigning** my job and planned to work in a **reputed** hospital. I was interviewed, excellent-performance and I got **appointment** letter with all the **salary** and **perquisites**. I was in an excited **mood**.*

Excerpt 6_49178115 (L1 Malayalam, FREQ3 15%, FREQAC 5%, SYNLE 2.93, ASL 9.07)

*The main difficulty which I faced was that he was from other religion. In our society religion has an important role to make decisions. Generally same religious people should marry each other. Otherwise people have no position in religious. According to me religious is not important than my love. So I preferred to marry him. I **informed** my views to my parents. My parents are*

moderate ***orthodox***. *They couldn't accept my views. But finally I got* ***permission*** *from my parents.*

The use of low-frequency vocabulary *(festival*, *tension, migrate, reputed, convince, orthodox, joy, blessing*) by these medium-proficiency second language writers is evident in these excerpts. The language is less well controlled than that of the high-proficiency group but the range of lexical resources demonstrated by the writers is still broad. Again, there is a wide variety of topics written on (careers, marriage, festivals) and they are all a direct response to the prompts. The non-restrictive nature of these prompts (open focus), in the sense that they allow writers a great deal of freedom in the topic of the response seems directly related to the lexical sophistication of the language used. The writer is not forced to produce prompt-specific vocabulary on a topic that he or she may be unfamiliar with but can choose a topic that is well suited to the lexical resources that the writer is able to draw on. This scenario is quite different for a prompt that does not allow for such a wide range of topics. More restrictive prompts, such as the one on child psychologists (prompt 95) force writers to respond to a topic that the writer may be unfamiliar with and have much less prompt specific vocabulary to draw on. This contrast in prompt characteristics (restrictive or non-restrictive) may well be related to the lexical sophistication of the responses.

6.3.1.3 Responses from the low-proficiency group

The following excerpts are drawn from the low-proficiency band of test takers who responded to prompts 73 and 115.

Excerpt 7 2595973 (Prompt 73, Writing score 70)

(L1 Tagalog, FREQ3 8%, FREQAC 4%, SYNLE 3.78, ASL 16.11)

I ***strongly*** *believed, that a person learned more from their mistakes than success. Cause realizing what you have done make you think not to do it again, instead, more on your life and make you better person or individual.*

There are lots of mistakes I once had in my life. Like when I was a student, particularly on my high school years. I remember, I didn't study my final ***exam****, and the outcome, I failed the subject. I was* ***relaxed*** *enjoying my* ***teenage*** *years. That my parents was really* ***upset*** *and kept on reminding me to study, otherwise, I'll be no body in the future. So then, it didn't take me long to realized that my parents are right. Thanks god, I finished my high school years on time.*

Later years of my life as a Nurse. There was once instances that I came in to work like 10 minutes like every Nursing Attendant were done with their report and I eventually miss the important part of my ***routine****, the* ***Endorsement****. If you missed the* ***endorsement*** *missed to know what the latest on your patient state of health, the* ***medication****, and everything. So, I start to get* ***panic****,* ***unstable****, I can't* ***concentrate*** *anymore. For then I learned to be on time now.*

On your road to success, you can't be successful without mistakes. A person who is success, that the person who know mistakes is what make you a good person

*In this result, **committing** mistakes can't be avoided, that sometimes we can't help, we are just human, and we are not perfect. As long as you REALIZED what you did and change for good will be successful in your career and life.*

Excerpt 8 49198115 (Prompt 115, Writing score 73)

(L1 Portuguese, FREQ3 11%, FREQAC 3%, SYNLE 3.59, ASL 13.69)

*My operation day was fixed on Oct 9th, 2001 as he was **breech**, low movements, the **placenta** was low level and early **contractions**. When they took the baby out, he **aspirated** **amniotic** **fluid** and became blue baby. Hope was less and he was in a critical situation. Immediately the concerned doctors **intubated** and moved to **ventilation**.*

Excerpt 9 49210115 (Prompt 115, Writing score 73)

(L1 Bulgarian, FREQ3 12%, FREQAC 3%, SYNLE 6.13, ASL, 21.2)

*So because of what happened my **grandmother** made the decision, that she will support me for my schooling but in one condition, that she will just give me 15 **cents** per day. When I was able to **recover** and stated me about her plans, I was so **excited** and willing to do whatever they will asked me to do like, cleaning the house, helping them to cook and doing the **laundry** too.*

The responses to these two prompts from the low-proficiency group are on a slightly more restricted range of topics than for the medium- and high-proficiency groups. There are more responses on the topics of the writers' professions (many healthcare professionals take the MELAB and that is apparent in the responses), school days, and weddings than from the other two proficiency groups. This trend perhaps reflects the more limited linguistic resources the low-proficiency group can call on.

The prompt categorization approach in this work classified prompts as open or focused. The intended distinction was between the contextualization provided in the prompt and the extent to which a writer was confined when composing a response (see p.61-62). Interestingly, although prompt 115 (memorable days) was categorized as focused, the test population has still responded on a range of quite different topics. That is, although the prompt was categorized as focused, the test population has not found the prompt to be at all restrictive.

Prompt 115 was categorized as focused because of the specific instructions given to the test takers in the form of the rhetorical cues. Although the cues do force the writer to include certain aspects of the story and specify some of the content of the response, the wording of the prompt still allows writers to select any personal experience that may be described as a memorable day. Hence, while the prompt may be focused, it is not restrictive in the sense that a specific topic is required in the response to the prompt, such that all test takers are compelled to write on the same topic. This difference in prompt characteristics appears to be more important in determining the textual features of the responses than the originally formulated distinction in focus. This reconsidered definition of the prompt characteristic of focus will be expanded upon in Section 8.1.2. Tasks that allow test takers to respond on a broad range of topics situated in the personal domain may be likely to elicit responses that are more lexically

sophisticated than those prompts that are situated in other domains and/or do not permit test takers to write on a broad range of topics.

The significant differences in lexical sophistication revealed by the MANOVA and post-hoc analyses (see Table 5.14 on p.99 and Figure 5.6 on p.100) are manifested in the lexical choices made by writers in the medium- and high-proficiency subgroups (and to a lesser extent those in the low-proficiency group). The fact that both prompt 73 and prompt 115 elicit narrative responses indicates that this is another characteristic that contributes to lexically sophisticated but syntactically simple responses with low levels of academic language use. The number of rhetorical cues is not a prompt characteristic that contributes to the differences in responses described above as prompts 73 and prompt 115 have quite different numbers of rhetorical cues.

It may be impossible to be conclusive regarding the prompt characteristics that contribute to high levels of lexical sophistication but the quantitative findings suggest that prompts situated in the personal domain, that elicit a narrative response, and that afford test takers the possibility of responding on a broad range of different topics (are non-restrictive) elicit responses that are significantly more lexically sophisticated than responses to other types of prompts.

6.3.2 Prompt 214

> How important is it to know about the personal life (e.g. health, personal relationships, youthful mistakes) of government leaders? What things should be made public? What things should be kept private? Give your opinion and support it with reasons.

Prompt 214 is situated in the public domain, elicits an argumentative response, is focused, and has a small number of rhetorical cues. The prompt elicited responses with the lowest levels of lexical sophistication. However, it elicited responses with longer sentences and longer words on average than most other prompts. In sum, the prompt elicits responses with lexically simple but syntactically complex written language.

6.3.2.1 Responses from the high-proficiency group

The following excerpts are from high-proficiency learners who responded to prompt 214.

Excerpt 10 49661214 (L1 Arabic, FREQ3 10%, ASL 19.76, AWL 4.39, Writing score 93)

> *Who hasn't heard of the senator who turned out to be having an affair, that member of **parliament** who – as an **adolescent** – smoked **weed** or even that candidate who is **accused** of sexual **harassment**? But how important are these stories for the public? Should the media really pay that much attention to the private lives of our politicians? In my opinion, this depends on the **gravity** of what that public figure is **accused** of having done, how **unethical** it is and how **harmful** it was to others.*
>
> *Monica L. Back then it was the story of the year, no newspaper could **resist** **dwelling** in it and almost everybody in the whole world had an opinion about it. But was it really worth it? Of*

*course, it was most **unfortunate** for his wife and daughter and I am sure he was very **embarrassed**. But putting these personal feelings aside, why was it the business of anybody else than him and his family? Did it make him a worse or better president? I don't think so and this example is a very typical case for me where private should stay private.*

*It's a **totally** different story though if we are talking about a sexual **harassment** **incident**, a situation where force has been used by the public figure or where somebody has been otherwise **harmed**. In cases like the ones just mentioned, we moved from a bad personal choice to an **ethical** issue that affects and **harms** others. I wouldn't want a president or member of **parliament** or any other political representative who can't draw a clear line between what's **ethical** and what isn't.*

*There are also known cases where some people who were running for office tried to use the fact that their **competitors** are suffering from a certain **illness**. If that disease isn't a serious **obstacle** that would keep the candidate from doing his / her job, then this should never be brought up neither by the **competitors** and **definitely** not by the media.*

*Generally, I don't think it is very difficult to **distinguish** between the things that the public is **entitled** to know and those things that are just used as **dirty** weapons against public figures. The question should always be: was the person's action **unethical**? Of course, **ethics** are very difficult to **generalize** and what's **ethical** for one person will be **unethical** for the other. But that would be a topic for another article.*

Excerpt 11 26901214 (L1 Punjabi, FREQ3 9%, ASL 16.85, AWL 4.72, Writing score 77)

*From individual's point of view, it is extremely important to know about the personal life of a government leaders. They are the persons who been the leader of the country, **functional** responsibilities to run the country. Election or **selection** of a wrong person will not damage the **reputation** of the **constituents** but will be quite **harmful** for the development of the country they belong to. If the leader has a good health he or she **devote** enough time **towards** **tendering** his or her responsibilities to the **constituents** in particular and to the country and rest of the nation in general. In case of **weak** health, no leader can perform a good job. Even if some is not good in health, one should not come forward to be a government leader.*

These responses from the high-proficiency group demonstrate good control over linguistic resources, typical of relatively proficient second language writers. The excerpts contain specific content that is a direct response to the prompt. What seems to distinguish these responses from the ones to prompts 73 and 115 is the specialized nature of the vocabulary that is required to respond to this prompt. These excerpts show that the writers from the high-proficiency group have written about the role that elected officials play in government, in addition to commenting on what voters need to know about them.

The language needed to discuss the role of elected officials is specialized and would be difficult to paraphrase on a timed test without access to a dictionary or other writing aids, as is the case with the MELAB. These demands on lexical resources are in contrast to those presented by prompts 73 and 115

where writers could address a broad range of topics of personal relevance. For prompt 214, if writers do not know the appropriate language to describe the functions of government and responsibilities of elected officials, it will be challenging to find high-frequency synonyms. Writers with less developed lexical resources are more likely to have to respond in generalizations and/or employ circumlocution, which may partially explain the low levels of low-frequency vocabulary used in responses to prompt 214.

6.3.2.2 Responses from the medium-proficiency group

The following excerpts are taken from responses to prompt 214 from medium-proficiency learners:

Excerpt 12 49711214 (L1 Thai, FREQ3 11%, ASL 30.2, AWL 4.67, Writing score 80)

> *Every human being in this has the right to maintain there personal life, no matter weather it is government leaders or **ordinary** people. People keep some **pleasure** while maintaining this **privacy**, but its all depends upon the **mentality** of a person political leaders life must be an open book for public, but **internally** some leaders may be not **satisfied** with this system.*
>
> *In the modern and democrat culture public think that they must know each and every aspect of the life of their representatives, because the political leader must be perfect in all levels of his life like, education health and public relationships. I think health is one of the major factor for the leaders, because without **adequate** health a person **cannot** work **properly** and he may make lots of **absence** during the week schedule. This will affect the public too much, because they may not get enough service **accordingly**. So the public have the right to know about the health status of their leader, then only they can try to make the changes of their government leader. To the other hand personal relationship and **privacy** with family is different, like the personal relationship with some companies and countries should be explained to the public in the other hand he can maintain his family life as a private matter. There is the next about **youthful** mistakes during the **adolescent** stage if a person create a bad record in his life, He is not **suppose** to keep the leadership position because a leader must be perfect than only his **followers** can get the **exact** directions. The public have the maximum right to know all about their leader.*
>
> *To put it in a **nutshell**, I can conclude that the government leaders life must be good model for his **co-workers** and for his fellow people.*

Excerpt 13 50569214 (L1 Arabic, FREQ3 5%, ASL 20.81, AWL 3.81, Writing score 75)

> ***Firstly**, the personal life of the leader is an important part of a leader's performance. For example, to most of **curious** people, knowing many details about their leader, like how many children he has, who is the lady he is married to, where did his parents come from or knowing about his relationships, is a common behavior that any leader should find it as a challenge to keep his personal life in a god condition. Therefore, these kind of personal information should not publish in public as a must, otherwise, the leaders themselves use them as a way to **advertise** themself.*

These excerpts provide examples of writers from the medium-proficiency group attempting to write lengthy sentences to connect together several supporting examples. The reason that this prompt elicited responses with significantly longer sentences than those of other prompts is not immediately

obvious but perhaps the specificity of the subject matter, combined with the argumentative responses contribute to the writers' attempts to connect ideas together in complex sentences. The second sentence of excerpt 13 above is an example of the test taker attempting a lengthy and complex construction, linking together multiple propositions in a single sentence. This type of sentence level writing is typical in responses to this prompt but much less common in the responses to prompts 73 and 115.

Also, the prompt provides test takers with some examples (e.g. health, personal relationships, youthful mistakes) of the types of supporting details that could be described. Test takers may take these examples and build them into the response, discussing one or more of the examples within a sentence. Possibly, the provision of examples within the prompt is a characteristic that may contribute to longer average sentence length. It would require further research to confirm whether this was indeed a prompt characteristic that contributes to the elicitation of different writing products.

Another possibility is that the argumentative response mode encourages writers to formulate longer, more complex sentences. The two prompts that elicited responses with the highest levels of syntactic complexity were both ones that elicited argumentative responses. Neither of these prompts were situated in the personal domain, which may contribute to the lack of lexical sophistication, but in contrast an attempt from writers to acknowledge the complexity of the subject matter addressed in the prompt by attempting to build a more complex argumentative structure, characterized by relatively long sentences.

6.3.2.3 Responses from the low-proficiency group

The following excerpts are taken from responses to prompt 214 from low-proficiency learners:

Excerpt 14 49638214 (L1 Punjabi, FREQ3 6%, ASL 14.38, AWL 4.21, Writing score 75)

> *How important it is to know the personal life of a political leader? It is important to know the personal background of a certain political leader* ***unable*** *to know how well he or she can lead the public he or she is been* ***elected*** *to. 1st a political leader should have a healthy way of living and mind, meaning in order to lead he or she should have a healthy body and a healthy relation with other people or* ***self****. It is important also to know everything in order to* ***correctly*** *chose the right one to* ***elect*** *to be a leader. If an individual is been* ***elected*** *with poor health will not be able to* ***fulfill*** *his or her duty and responsibility and will not attain his or her full capability as a leader, it will be always a* ***hindrance*** *on his or her success as a leader. Personal life is a factor to know in choosing a good leader, one should know ones personal life. Knowing a personal life give good picture in how he or she treat his or her family. If an individual treat his or her family* ***poorly*** *just imagine how she or he will rule or lead. If an individual run for a position it means the she or he* ***surrender*** *all his or her personal life that's the price of becoming a leader. No secret should be keep everything should be* ***transparent****. In real world its very hard to know which the one to be sitted as leader. The only thing that learned in this life is that we will only now the truth when we already make the decision. As for me regarding this* ***topic*** *in pursuing a sit on a political*

bandwagon *one should* ***disclose*** *everything, one should be honest and open about it. Cause hiding things especially from the public is like* ***cheating*** *them face to face.*

Excerpt 15 50563214 (L1 Arabic, FREQ3 7%, ASL 13.36, AWL, 3.90, Writing score 65)

People in the world want to live their life. They always want to live happy, but that is impossible because there are somethings are good which make us happy and there are other things make us ***sad****. To everyone around the world lives in* ***relax*** *life, the countries must avoid a lot of things. One of the most important things is the personal life. However, how people can live without personal life.*

While both these writers had difficulty expressing their ideas, there is still evidence of attempts to construct relatively complex sentences. The second sentences of excerpts 14 and 15 are examples where the writers have attempted to connect several ideas together into relatively complex constructions. While the attempts are not successful, they can still be processed for meaning by the reader and they are indicative of the tendency in responses to this prompt to construct relatively long sentences.

Test takers in the low-proficiency group are unlikely to have acquired the relatively specialized and low-frequency vocabulary needed to describe functions of government and are left to focus their more generic responses on what kind of person is suitable to be a public leader. Hence, the responses become more descriptive in nature, in contrast to the more sophisticated arguments presented by the higher proficiency test takers.

6.3.3 Prompt 108

In some countries such as the United States, the standard work week is 5 days, 8 hours a day. Some companies are considering changing their work week to 4 days, 10 hours a day. What are the advantages and disadvantages of the two options? Which do you think would be better for society? Give reasons to support your opinion.

Prompt 108 is situated in the occupational domain, is focused, elicits an argumentative response, and has a large number of rhetorical cues. Prompt 108 elicited responses with the longest average sentence length and the highest values of syntactic left embeddedness, the two discourse measures that operationalize syntactic complexity (see Figure 5.8 on p.104). This prompt elicited the most syntactically complex language of all the prompts in the study. Prompt 108 also elicits responses with low levels of lexical sophistication. It elicits the second lowest level of low-frequency vocabulary after prompt 214 (see Table 5.18 on p.108).

6.3.3.1 Responses from the high-proficiency group

The following excerpts are taken from responses to prompt 108 from high-proficiency learners:

Excerpt 16 45243108 (L1 Nepali, ASL 31.67, SYNLE 6.80, FREQ3 5%, Writing score 87)

*"Work is life and life is work." We human beings can agree with the statement because all the basic as well as the **sub-ordinate** needs of every human being can be **fulfilled** only when he or she works to earn for his or her life. People need money to **fulfill** their needs. So they need to work at some field like business, administration, teaching, medicine, industry, **agriculture**, etc. In many developed countries it has been found that people work hard to earn good money for their family and themselves and usually the working hours are eight hours a day and five days a week. They get two days holidays in a week as weekend. Some companies may have working hours as ten hours a day and four days a week having three days holidays. Whatever the total working hours in a week is forty hours.*

*When people are working five days a week and eight hours a day, they can be benefited by being able to provide **sufficient** time to their family daily in the morning and evening in the work days as well as they've got two days weekend holidays when they can do **laundry**, cleaning and gardening as well. But sometimes it can be **disadvantageous** because people may want long weekend for completing all their **household** tasks and they also may want to go out for a short **vacation**. For that purpose if people are working four days a week and ten hours a day, they can get three days long weekend and can spend lots of time with their families and friends as well so that people can be more important for their society. They start to become more social and the isolated nature of people may go on reducing **tendency**.*

*When people are working ten hours a day they may not be able to give more time daily to their family in the mornings and evenings and it may create too much stress on every family members for all those four days in a week and even if they get three days long holidays or weekends they may not want to go out on **vacation** or friends place, they may want to be more inside home.*

*In conclusion it can be said that since the working hours is the same i.e. 40 hours a week, it can be always **beneficial** or **suitable** for people to have work life balance when they work eight hours a day and five days a week having two days holidays people may not be able to complete all the official or government office related works in the four working days if there are three days government holidays in a week. In my opinion people can be the **valuable** members of a society when they are working five days a week and eight hours a day.*

Excerpt 17 48208108 (L1 Tagalog, ASL 28.84, SYNLE 4.42, FREQ3 9%, Writing score 80)

*Having a 5 days 8 hours work days have **disadvantages** and advantages. The **disadvantages** are as follows: one, people trend to work longer days which is 5 days and only gets 2 days off. Sometimes 2 days off are not enough to be able to get to do the things you want to be done on four days off, example, you want to take more sleep hours on the days you don't have work, do the **laundry**, do the **errands** of just **relax** watching TV with your family or friends, if not by yourself. Second, it feels like 2 days are only doing the **errands** you were not able to do on the **weekdays**, like do the shopping and clean the house!*

The long and complex sentences within these excerpts are quite evident. There seem to be at least two contributing factors to the length of the sentences in responses to this prompt. First, some of the prompt wording (5 days, 8 hours a day; 4 days, 10 hours a day) is regularly repeated within the response and this repetition of propositions from the prompt adds several words to some sentences. Second, the supporting examples that are produced by writers in response to this prompt are quite simple (for example, doing household chores, visiting family and friends, taking a vacation), both cognitively and linguistically. When it is both cognitively and linguistically simple for test takers to produce many supporting examples, they can write long strings of these supporting examples, which greatly adds to the average sentence length. Also, prompts that elicit an argumentative response that is not in the personal domain tend to elicit responses with relatively long and complex sentences. This seems to be the case with prompt 108 as it was with prompt 214.

6.3.3.2 Responses from the medium-proficiency group

The following excerpts are taken from responses to prompt 108 from medium-proficiency learners:

Excerpt 18 48014108 (L1 Mandarin, ASL 25.22, SYNLE 5.89. FREQ3 8%, Writing score 73)

> *As the developed countries, such as United States, Canada, or European countries, considering changing working time, in my opinion, has both advantages and **disadvantages** on the aspects of company **productivities**, family activities.*
>
> *The changing working time brings much influence on company business especially for the **manufacturers**. The standard work week is 5 days, 8 hours a day has provided the public more service than less work week in 4 day, 10 hours, for example, people need to go to banks or governments for specific tasks on Friday. As such, some important thing happened can't wait over the next week work time if the work time in government is **shorten**. **Nevertheless**, some will say if the working hours **prolongs** two hours every day, but only 4 days in a week, they can take the **flexible** jobs and go to handle their banking matters after work, as the government open the service 2 hours longer than before. In addition, **shortening** the working days results in the **productivity** of **manufacturing** schedule. The **manufacturers** will produce less products and cost will increase because the machines **shall** stop for 3 days as no workers on duties.*
>
> *Another effects of changing working week is for family activities. Families likes to have more time spending with family members, children together; however, if school only operates for 4 days, some courses can't finished.*

Excerpt 19 48028108 (L1 Farsi, ASL 18.47, SYNLE 4.40, FREQ3 13%, Writing score 80)

> *Some people believe that, as I do, there should be 5 days per week, 8 hours a day because in this way they have more days to work and at the same tie especially when they are working with some countries such as Arabic countries in the middle east, e.g; U.A.E. where the weekend is Friday so they will have 3 days **gap** which is Saturday, Sunday and Friday. In addition they find it bearing to work 10 hours a day, because they believe that the work **efficiency** will be **decreased***

after 8 hours. They say the human being needs to take rest after 8 hours of work. I think this way of working, 5 days per week 10 hours per day is desired by most of the employees who are so busy in their business and have international business.

The use of time phrases from the prompt is again evident within these excerpts, which contributes to some long sentences. The second sentence of the second paragraph of excerpt 18 is a clear example of repetitious use of propositions from the prompt contributing to long sentences within the response.

6.3.3.3 Response from the low-proficiency group

The following excerpts are taken from responses to prompt 108 from low-proficiency learners:

Excerpt 20 48155108 (L1 Arabic, ASL 27.82, SYNLE 8.46, FREQ3 9%, Writing score 73)

*Many people would like to work eight hours a day, and some people prefer to work more than that. Such as ten hour a day. So people who prefer to work week to four days, teen hours a day, they have more options than people who work week to five days, eight hours a day because they have more time to do their **chores**, they can **communicate** with friends and family, and they have more time to enjoy their life.*

*First of all, people who work five days a week, eights hours a day, they can not do their **homework** as the person who work four days a week, ten hours a day because they will just have two days off each week; however, people who work four days a week, they have three day off each week. So they can make time management, and they can organize their time to do their home work such as **laundry**, cooking, and clean their house.*

*In addition, they also have more time to **communicate** with their friends and family. **Nevertheless**, people who work five days, it will be hard for them to contact with their friends and family since they do not have more time in the week. On the other hand, it is a very significant for people to have more time to **communicate** with other because they should have to have more time each week. For instance, some people do not know their family because they have to work a lot, and they just focus on their work, and they do not care about their family and friends.*

Finally, people should take a rest and enjoy their life, however, some people need to have more time, but they can not because their company recommend them to work a lot.

*In short, **chores**, communication, and **enjoyment**, they*

Excerpt 21 48002108 (L1 Punjabi, ASL 21.15, SYNLE, 7.0. FREQ3 7%, Writing score 75)

*First, if a individual is working eight hour day and he/she have enough time to look after his or her family, after working eight our a individual can be involved in their kids's live after school activities for example **soccer** game or any other game. Being involved in his/her kids's life their*

interpersonal *relationships get better and it helps the kids to be open communication with their parents. When these children go to high school they make better decisions compared to those children who do not have good* ***interpersonal*** *relationships. Working eight hour and five days a week individual always have time on weekend and they can make plans as a family trip somewhere.*

The first sentence of the second paragraph of excerpt 20 provides a very clear example of a test taker using the language from the prompt to construct a very long sentence. The structure of the sentence is not particularly complex but the joining together of the propositions from the prompt allows the writer to produce a relatively complex sentence.

The prompt characteristics that contribute to the long sentences for prompt 108 appear to be the particular phrasings or propositions within the prompt and the topic that allows for cognitively and linguistically simple supporting examples to be produced at length. In addition, the response mode (argumentative) for both prompts 214 and 108 seem to contribute to the high levels of syntactic complexity in the responses to both prompts. The low levels of lexical sophistication in responses to prompts 214 and 108 are likely attributable to the prompts not being situated in the personal domain and not allowing test takers to respond to a broad range of topics.

6.3.4 Prompt 95

Some child psychologists believe that the peer groups children play with influence their character and personality development more than the children's parents do. The psychologists say children are more interested in fitting in with their friends than behaving the way their parents want them to. Do you agree or disagree with these psychologists? Explain your point of view.

The one other prompt (after 73, 115, 214, and 108) that stands out as eliciting responses with significantly different written products is prompt 95. Prompt 95 is situated in the public domain, is focused, elicits an argumentative response, and has a small number of rhetorical cues. It elicits responses with significantly higher levels of academic vocabulary use than other prompts (see Figure 5.7 on p. 102). 8.25% of the responses, on average are made up of vocabulary that is classified as academic in the COCA corpus (see Table 5.20 on p. 110). The next highest value for the percentage of academic language use was 6.23% for prompt 100. Prompt 95 also elicited responses with the longest average word length. Please note that in these samples, words classified as academic in the COCA corpus are bolded and underlined.

6.3.4.1 Responses from the high-proficiency group

The following excerpts are taken from responses to prompt 95 from high-proficiency learners.

Excerpt 22 4711695 (L1 Farsi, FREQAC 21%, AWL 4.92, Writing score 85)

> *Human personality **development** is one of the most important and most studied fields in **psychology**. From earliest time of studying child psychology, psychologist have been interested in explaining the **factors affecting development** of a child. There are many **theories** and very **well-documented research** in this **topic**.*
>
> *I believe that the **development** of child personality is such a **complex** and **multidimensional phenomenon** that cannot be explained by the **role** of one **factor** such as **role** of **peer** group playing or the **role** of parents. To my view, instead of arguing in favor of one **viewpoint** and **disregarding** the other one we have to **study** this intricate and **multifaceted characteristic** by **analyzing** all **factors affecting** its **development** and **evolution**.*
>
> *even though there is a strong agreement on the **importance** of the **role** of **peer** group playing in child's personality **development** but its better to **approach** this argument by considering the **stages** of life of a child and to **determine** what **factors** are more **influential**. For **example** in the first year of child's life this is only the mother who **affects** the most. in the second stage both parents have some **roles** in the personality **development** and later the other siblings play some **role** in child's **development**, **however** when the childs arrives to school age the **teacher** and other kids have important **role** in child's personality. **Moreover**, in **adolescence** we can expect the most role played by **peer** groups in*

Excerpt 23 4712995 (L1 Tagalog, FREQAC 15%, AWL 4.83, Writing score 77)

> *Tell me who your friends are and I will tell you who you are." A statement that may support a **study conducted** child **psychologists stating** that children's **behavior** and personality are more **influenced** by peers than their parents. Child **psychologists** are people who specialize on how people acts and behaves focusing on children. They **initiated** this **study** considering **factors** that may influenced a child personality, which they **concluded** that it is friends who plays a major part.*

The high proportion of academic vocabulary is clear in both these excerpts, however some of the vocabulary categorized as academic has been lifted from the prompt wording. The topic presented in the prompt is an invitation to produce field-specific terms if a test taker knows them. In the case of these high-proficiency test takers, they have been able to demonstrate their ability to appropriately use some field-specific terms (despite some having been lifted from the prompt) and other academic language in direct response to the prompt.

6.3.4.2 Responses from the medium-proficiency group

The following excerpts are taken from responses to prompt 95 from medium-proficiency learners.

Excerpt 24 4706395 (L1 Tagalog, FREQAC 17%, AWL 4.87, Writing score 73)

*The children play a huge **influence** on its character by **engaging** to its **peer** rather than to his/her parent because I believe that children can change character by other children such as **demonstrating** **competence**, **influencing** confidence and freedom in expressing feeling.*

*Children can **demonstrate** their **competence** to other by **participating** **activities** in school, for instance, in a **examination** a **student** won't be on top before other. In their way they showed their **competence**. A child can **acquire** this kind of **behavior** to **peer** **grouping** because they want to be more competitive to its **peer**.*

*A confidence of a child start to their parent but **peer** **grouping** can boost **peer** confidence. For **example**, a child that is belong to a family that is ashamed of immersing to other people, **peer** **grouping** boost confidence by giving the child a responsibility to socialize to other people, and a child can **acquire** **self-esteem** to **peer** **grouping**.*

***Lastly**, the freedom of expressing feeling can **motivate** by **engaging** and **grouping** to other children. In this condition, a child has the freedom to speak whatever can come to its mind; in **particular** a child has a problem to his/her parent, this child can express it feeling to other child.*

*In **conclusion**, children is more comfortable in **peer** **grouping** rather than parent do. Because, they feel better to be with their friend and more confident to face their **community**. **Peer** **grouping** gave the child to boost their **self-esteem** and a **positive** **behaviors** to the **community**.*

Excerpt 25 4708495 (L1 Gujarati, FREQAC 12%, AWL 4.63, Writing score 77)

*In **addition** **socially** parents are upbringing their children with **religious** group **community** **services** and more. So that they know how to mingle with people. When I was a kid my parents will get us to **social** gatherings and there by me and my sisters know more about **culture**, **values**, accepted **behavior** system in the **society** and more.*

The test takers who produced these excerpts have less control over their lexical resources than those from the high-proficiency group. Some terms are used incorrectly (peer grouping), while other terms are used appropriately but the writer lacks the ability to use the terms in a well-controlled sentence, such as the first sentence of excerpt 24. However, these two test takers were still able to demonstrate knowledge of a relatively large amount of academic language, they just are less able to control it appropriately than the higher proficiency test takers. Prompt 95 exerts a similar effect on these writers as it does on higher proficiency language learners, indicating that the characteristics of this prompt that contribute to differential performance (high levels of academic vocabulary use) affect not only highly proficient learners. The main characteristic of this prompt that contributes to significantly different written products seems to be the academic subject matter.

6.3.4.3 Responses from the low-proficiency group

The following excerpts are taken from responses to prompt 95 from low-proficiency learners.

Excerpt 26 5046295 (L1 Malayalam FREQAC 12%, AWL 4.97, Writing score 70)

*"Today's children are tomorrow's citizens". Children are fun loving, **innovative** and agile. They are more interested to make friends. In my point of view, the **peer** group **influence** the children for their character and personality **development**. Parents are the first **teachers** of their children. Anyway parents can teach good things from home in their early **stages**.*

***Firstly**, I would like to **substantiate** my views. From school children will get a **peer** group. They try to imitate the other children. **Teachers** also plays an important **role** to mold a child's character and **behaviour**. In earlier stage children only imitate their friends. They will learn everything from their friends.*

*Secondly, parents can also play an important **role** to mold a character of child. They can teach their children nicely. From home they start to learn things. **Education** makes a child to **develop** good **behaviour**. **Education** **liberates**. **Education** **enlightens** people and **empowers** them.*

*In some schools **provide** personality **development** **classes**. It is good for the children. Children are more comfortable in the **peer** group. They try to grasp everything from their friends. "One rotten apple spoils a basket of apples". I like to agree with this. Some children with misbehaviour spoils every children from his group some bad **behaviour** **develops** according to the **influence** of bad **students** from the early states they can develop this behaviour. Some **psychology** says children are more fitting with their friends.*

*To put it in a nut shell, children are more **influenced** by their **peer** group. In any age group they try to imitate their friends and grasp ideas from their **peer** group.*

Excerpt 27 (L1 Tagalog, FREQAC 12%, AWL 4.66, Writing score 73)

*In **conclusion**, aside from having **experience** their parents company, children in **peer** **groups** are not just contented. They **develop** their **knowledge** from **peer** **groups** that **influences** the character and personality **development** more that the children's parents do. If I will be lucky to have a child I will protect **him/her** for bad **peer** **groups**!*

Even these low-proficiency learners are able to produce relatively large amounts of academic language in response to this prompt. The choice of verbs following "education" in the third paragraph of excerpt 26 is particularly impressive from a low-proficiency test taker. Excerpt 27 relies on language from the prompt for much of the vocabulary that is categorized as academic but is able to do so in a way that produces a comprehensible paragraph. While the lifting of academic vocabulary from the prompt wording contributes to the high rate of academic vocabulary use in these responses, many test takers at all proficiency levels have demonstrated their ability to take this academic vocabulary and use it to produce aa coherent response to the prompt. There certainly appears to be some feature of this prompt that elicits high levels of academic language from test takers regardless of their proficiency level. The academic subject matter of the prompt topic is likely the contributing factor to the high levels of academic language found in these responses.

The fact that this prompt elicited such high levels of academic vocabulary use, relative to the other prompts studied is perhaps indicative that such prompts (those posited on very specific and academic topics) can elicit responses that are significantly different, in terms of the final written product than responses to other prompts. This suggests that to help establish task equivalence, prompts with such specific academic subject matter should not be presented on the same test as prompts with more generic, non-academic subjects.

6.4 Summary of findings

The excerpts given in this chapter are drawn from responses to the prompts that elicited the most significant differences in textual features (73, 115, 204, 108, and 95). These excerpts provide contextual evidence of the differences in the written products, produced in response to different prompts. The excerpts from responses to prompts 73 and 115 (prompts situated in the personal domain, with an open focus, and that elicit narrative responses) were characterized by high levels of low-frequency vocabulary, compared to the responses to other prompts. The responses to these prompts were also written on a broad range of different topics, indicating that the open focus of the prompt has an effect on the textual features of the response. Both these prompts elicited narrative responses, which indicates that the response mode is also a prompt characteristic that has an effect on the textual features of the response. The interactions between the three prompt characteristics (domain, focus, response mode) that characterize prompts 73 and 115 are difficult to untangle but the data and a reading of the responses indicate that all three prompt characteristics should be controlled for if prompt equivalence is to be established. Further discussion of these findings will be provided in Chapter 8, particularly in Sections 8.1.1, 8.1.2, and 8.2.1.

Prompts that elicited argumentative responses were characterized by syntactically complex writing. There was not a significant difference in the overall length of the responses (total number of words in the response) across response modes but the argumentative responses featured longer sentences than the narratives. Interestingly, the attempt to construct lengthy sentences in argumentative responses was a feature at all proficiency levels. Even test takers from the low-proficiency group attempted to write sentences that connected a number of ideas together. The resulting sentences were frequently inaccurate but the response mode prompt characteristic has an effect on the textual features of the responses for both narrative and argumentative responses. These textual features (complexity of sentence structures and connection of ideas) are prominent within the MELAB Writing rating scale descriptors, indicating that these important indicators of second language writing proficiency are captured on the MELAB Writing Test.

Other prompt characteristics that had an effect on the written products were specific phrasings or propositions within the prompt wording and the use of academic subject matter in the prompt. These two characteristics had a significant effect on the syntactic complexity and the level of academic language use, respectively within the responses. Prompt that features specific propositions or phrasings that must be repeated over and over within the responses (as in prompt 108) should be avoided. They inflate the average sentence length of the responses, which could arguably be seen as an indicator of

syntactically complex writing that is actually an artefact of the prompt rather than the ability of the writer. It is a feature that has an effect on the textual features of the responses and should be avoided during the development of the prompt. Whether overtly academic subject matter is permissible on a test program will depend on the construct that is being measured. What was apparent from an analysis of the responses to prompt 95 was that the topic of the prompt (children's social development) was an invitation to use as much academic vocabulary as the test takers were able to produce. The responses to this prompt featured significantly higher rates of academic language use than did responses to other prompts and this difference is attributable to the prompt topic. The prompt directly asks test takers to write on the topic of child psychology, parental influence, and peer group effects. The prompt also requires the test taker argue a specific opinion about the prompt topic. These features of the prompt contribute to the very high levels of academic language use in the responses.

Overall, the reading of the responses to the prompts described above helped clarify some of the textual features that differed significantly across prompts. It helped reinforce the fact that several prompt characteristics (domain, response mode, and focus) have an effect on the responses that is detectable statistically but may also be observed by the reader.

6.5 Implications

The findings reported above paint a varied and rich picture of prompt effect on a high-stakes writing assessment. On one hand, the written responses elicited from different prompt characteristics vary significantly by a variety of key textual features. The lexical sophistication, use of academic vocabulary, and syntactic complexity of the responses differ significantly depending on the characteristics of the writing prompt. This is a strong indication that the prompts studied in this work cannot be considered equivalent as they do not elicit written products that are consistent or comparable in terms of key textual features. Responses to prompts 73 (mistakes) and 115 (memorable days) that are characterized by narrative responses in the personal domain do not seem to be consistent with the test purpose of the MELAB (see 4.2.1): to assess the readiness of nonnative speakers of English for admission to college or university or for professional licensure purposes. Although the construct assessed by the MELAB Writing test is not clearly defined, narrative responses in the personal domain do not seem to provide very relevant evidence to the ability of the test taker to write in the academic or occupational domains.

However, while there is clear evidence that some textual features of the response will vary depending on the prompt, the quantitative analyses also showed that the holistic score awarded to the responses will not vary significantly based on the prompt. While it is reassuring that the test takers' scores do not vary as a result of the prompt it is perhaps surprising that the scores are stable considering the differences in textual features within those responses to different prompts. What does it say about the meaningfulness of the scores if a lexically sophisticated but syntactically simple response may be awarded exactly the same score as another response that has the opposite textual features? In addition to the inferences that may be drawn from the writing test scores, the findings also raise questions about the rating scale and the operationalization of the scale by the raters. As scores do not vary even when responses differ by key textual features, the scale may be insufficiently clear for raters so they may be unable to award a score that distinguishes between responses that are very different in terms of their

textual features. The holistic ratings awarded to MELAB Writing responses may partially explain that fact that scores did not vary significantly. Although the textual features that differed significantly are addressed in the MELAB Writing rating scale, the fact that raters award only a single holistic score may partly obscure the differences in the written responses. Analytic scores that were directly relevant to the textual features that exhibited significant differences (syntactic complexity, lexical sophistication. Academic vocabulary use) may have demonstrated significant differences where holistic scores do not. Alternatively, the raters may not be sufficiently well trained to be able to interpret the scale and to apply it consistently.

The textual features that most contribute to the significant differences in written product; syntactic complexity, lexical sophistication, and academic vocabulary use, are not all addressed within the MELAB Writing rating scale. Academic vocabulary use is the one textual feature that is not specifically addressed in the scale. The scale requires raters to evaluate the range of vocabulary used in responses at almost all score points but the distinction between academic vocabulary use and general vocabulary use is not explicated for the rater. This is a potentially valuable distinction to be captured in the scale, based on the findings of this study and could be a useful revision to the scale that will better allow important differences in written products to be distinguished.

There will be further discussion of the implications for inferences that may be drawn from the scores in Chapter 9 (see p.175). The potential implications of the findings for the rating scale and the behavior of the raters are also quite serious. If the same scores may be awarded to different responses for different reasons, there is a possibility that the rating scale is not able to capture differences by key scoring criteria (vocabulary and complexity).

The MELAB Writing rating scale is presented in full in Appendix 2 and it can be seen that explicit reference to use of syntactic structures and use of vocabulary is made at almost every score point. However, as a holistic scale, only a single score may be awarded by a rater to an essay. If a response demonstrates vocabulary use at one score band but syntactic control at a quite different band, the rater must make a determination which of these traits wins out. As only a single score may be awarded, responses with uneven profiles, in terms of vocabulary use and syntactic control must be assigned a score that better reflects one of those traits over the other. This is a perennial concern of holistic scoring and the findings of this work bring into question the relevance of using a holistic scale to evaluate responses to prompts with such varied prompt characteristics. In the case where prompts with substantively different characteristics are administered across test forms that must be parallel, there is a strong case for scoring written responses analytically, so responses with uneven textual profiles may be awarded the trait specific scores they deserve by, for example awarding one score for vocabulary use and a different score for syntactic complexity.

Finally, even with a rating scale that is well designed, the training and monitoring of raters needs to be effective to ensure that raters consistently apply the scale appropriately. If a test program administers prompts with a range of characteristics that may vary across test administrations, the raters will need to be trained how to apply the scale consistently to responses with quite different textual profiles. The rating will need to demonstrate to raters how to score narrative responses alongside argumentative

responses, if those prompt types are permitted on a single test program. Raters will then need to be monitored to verify that they are scoring responses in a manner that is reliable. The greater variety there is within a test program, in terms of the range of permitted prompt characteristics, the more challenging it will be to train and monitor raters when scoring responses that vary by several key textual features.

In summary, the quantitative analyses reported in Chapter 5 and the review of written responses in this chapter have confirmed that the prompt characteristics studied in this work have a significant effect on a range of textual features within the responses to several different prompts. The responses to the different prompts are not stable or consistent in terms of the textual features of the responses. This key finding brings into doubt the equivalence of the writing prompts studied and suggests that these prompts do not provide an equal challenge and opportunity to all test takers. However, although there are observed significant differences in several textual features, the accuracy of the responses does not vary by prompt and the score awarded to the prompts is also not prompt dependent. These are the findings and implications that may be reported based on the quantitative approach in this study. The findings and implications will be revisited and discussed further in Chapters 8 and 9 after the qualitative approach within this work has been presented in Chapter 7.

Chapter 7 – Qualitative study: materials, methods, and results

Chapters 5 and 6 focused on how the characteristics of writing prompts affect written products. The findings indicated that several prompt characteristics affect the nature of the written language elicited. The lexical sophistication, level of academic language use, and syntactic complexity of responses varied significantly according to the prompt. These findings were derived from a quantitative approach to investigating prompt effect. Similar quantitative approaches have dominated the second language writing assessment field (see the review of the prompt effect literature in Chapter 2, from p.21). Although quantitative methods predominate in the literature, relationships between prompts and written products depend upon the choices made by a writer or test taker in selecting and then responding to a particular prompt. The decisions made and actions taken by the individual writer are central to the effect the prompt characteristics have on the written product, but these are not well captured by quantitative methods. Qualitative research into decisions made by test takers may complement quantitative findings by providing insights into how test takers interact with and respond to prompts.

A series of interviews was conducted with test takers at a major *Michigan English Language Assessment Battery* (MELAB) test center. The aim was to explore how test takers selected a prompt and how they utilized that prompt throughout the writing test. The data collected from the interviews helped to answer the third research question addressed by this work.

How do the writing prompt characteristics affect the test takers' test taking processes?

This chapter will report on both the methodology employed and the results derived from the investigation into how writing prompt characteristics affect test taking processes.

7.1 Methodology

MELAB test takers are presented with two independent writing prompts. They must select one of those prompts and have 30 minutes to compose an essay. Interviewing the test takers after they had completed the test gave them an opportunity to recount how they had decided which prompt to respond to and how they had interacted with that prompt throughout the writing test.

The few qualitative studies of second language writing assessment undertaken (see Sections 2.4.2 and 2.4.3), consisted of interviews with test takers. These studies were reviewed in Chapter 2 and showed that high-stakes assessments could have a negative effect on students who felt they were unfamiliar with the linguistic or cultural demands of the test. However, the aims of this current work are somewhat different from those of those previous studies. Those studies investigated the broader test environment and its impact on the test takers, whereas this study is focused specifically on how the characteristics of the writing prompt influence test taking processes.

To understand the thought processes test takers undergo as they select and respond to a writing prompt, a form of verbal protocol analysis (see, for example Ericsson and Simon 1980, 1985, 1987) called stimulated recall was adopted (Gass and Mackey, 2000; Greene and Higgins 1994). The aim of the stimulated recall interviews was to capture, as fully as possible the thought processes (decisions and resulting actions) that the test takers went through as they completed the writing test. The stimulated recall interviews followed the guidance provided in Greene and Higgins (1994: 123) and were recorded for later transcription and analysis.

Greene and Higgins provide some detailed guidance for the collection of retrospective protocol data. The term *retrospective protocols* is used to distinguish the approach from that of *concurrent* (or thinkaloud) protocols that require participants to describe what it is that they are doing while they perform the actions under observation. As the actions to be observed take place during the live administration of a high-stakes test and because concurrent protocols would necessarily impact on how the test was administered, retrospective protocols were preferred. Beyond these practical constraints, previous qualitative research into writing processes (Rose, 1980, 1984; DiPardo, 1994) has shown the benefit of the retrospective approach over concurrent verbal protocols, which can interfere with writers' cognitive processes while writing.

The following guidance provided by Greene and Higgins (1994) was employed for the retrospective protocols performed with MELAB test takers.

- Collect data immediately after the performance

Data needs to be collected as soon after the event to be recalled as possible. This is due to the constraints of working memory (Ericsson & Simon, 1980; Garner, 1982). Minimizing the length of time between the event and the report is essential for a thorough recall of the decisions and thought processes undertaken.

- Focus on critical incidents and contextual cues

Retrospective reports can tend to elicit general accounts of the event of interest rather than specific recollections. However, when "researchers ask writers to reflect on concrete examples of writing, rather than writing in general, they are more likely to obtain more detailed information," (Greene & Higgins, 1994: 123). This is why the interviews with the MELAB test takers used the interviewee's written response to the MELAB writing prompt in the interview. The MELAB writing section is completed by hand so the answer booklet, with the test taker's handwritten essay was taken into the interviews so the interviewee could have a concrete example of writing to reflect on. The essay was used to stimulate the recall of the time when the writer was actually thinking, planning, or composing.

- Clarify the purpose of retrospective accounts

Greene and Higgins (1994) underlined the need to make the aims of the interview and the role of the interviewer clear to the interviewees. Both the purpose of the interviews and the role of the interviewer were made clear orally and in writing before the interviews began, as will be explained in more detail in Section 7.2.

- Design clear prompts

Referring to the line of questioning used in the interviews, rather than any actual test prompt or task, questions should be based on events that have already occurred and should not be hypothetical (Greene and Higgins, 1994; Nisbett and Wilson, 1977). Leading questions must be avoided and open-ended questions should be favored. Questions should move from general to specific. "We can pick up on interesting threads that emerge in students' explanations and we can ask students to elaborate on these as they emerge (Greene and Higgins, 1994: 126). This is the approach that was adopted for the interviews with MELAB test takers and will be described in detail in Section 7.2.3.

- Use converging methods

Greene and Higgins (1994) emphasized the need to analyze the text produced by writers in addition to the information that can be elicited form the interview data. They cautioned that conclusions drawn from only interview data may be misleading and partial. The retrospective protocol data plays a supporting role in this research to the quantitative analyses of the written products.

7.2 The interview process

A series of interviews was conducted between the researcher and 28 MELAB test takers at the Toronto MELAB test center. The interviews were done in two phases; an initial pilot phase when ten test takers were interviewed and a second phase of an additional 18 interviews when the data that would be analyzed for reporting was collected. The purpose of the first phase of interviews was to refine the questions and techniques that would be utilized in the second, or main data collection phase.

Table 7.1: Phases of stimulated recall interviews

Phase	Location	# of interviews	Purpose
1	Toronto MELAB Center, University of Toronto, Canada	10	Pilot
2	Toronto MELAB Center, University of Toronto, Canada	18	Main

All interviews, for both phases were conducted following the end of a MELAB test administration (after the writing, listening, and reading sections), meaning that the test takers were being asked to reflect on their experience of completing the writing test within two or three hours of having completed it. It was not possible to interview the test takers any sooner after the writing test, without interrupting the test administration.

7.2.1 Participant recruitment

At the end of the writing section, test takers were asked whether they would be willing to participate in a research project into the MELAB writing test. This invitation was extended orally by Toronto MELAB test center staff and not by the researcher. Test center staff asked whether test takers would be willing

to talk about their experiences of taking the writing section and that about ten minutes would be needed.

It was made clear by the test center staff that this was a voluntary opportunity and that no test taker was required to participate. They also emphasized that taking part in the interviews would in no way influence the volunteers' scores on the test. Anyone who was interested in participating in the interviews was asked to stay behind after the administration of the MELAB was completed. During Phase 1 of the interviews, the MELAB test center staff introduced me as an employee of the University of Michigan (the institution that makes the MELAB). This proved to be problematic during some of the interviews, as some interviewees wanted to discuss how to prepare for the MELAB and/or how they could do better on the writing section. In an attempt to address this issue, during Phase 2 of the stimulated recall interviews, I was introduced to the MELAB test takers as an independent researcher from a UK university. This approach to the interviews successfully eliminated the distracting questions about test preparation. This approach was reviewed at the Change of Status upgrade when this work was cleared for completion as a doctoral degree at The University of Nottingham (see Appendix 5).

Each test taker was interviewed individually by the researcher in private. All interviewees signed an informed consent form (see sample in Appendix 5) before beginning the interview process. The informed consent form described the purpose of the research and the aims of the interview. It also stated that interviewees may withdraw their willingness to participate in the process at any time. All interviews were digitally recorded with the permission of the interviewee. As the interviews were conducted between only the interviewer and a single interviewee, the recording was straightforward and followed the guidance (placement of recording device, proximity of interlocutors, need for backup equipment) provided by Leander and Prior (2004).

7.2.2 Pilot of stimulated recall interviews

The two interview sessions that made up the pilot phase were conducted in the first half of 2010 at the Toronto MELAB test center. Ten test takers were interviewed in total. Table 7.2 shows the writing prompts that were administered on the dates the test takers were interviewed.

Table 7.2: Prompts used in pilot (Phase 1) of stimulated recall interviews

Date	Number of interviewees	Prompts
3/20/2010	5	Some businesses tend to employ workers who are almost all in the same age group. Other businesses have a broad range of ages in their workplaces. What do you think different generations contribute to a workplace, and what effects do you think there might be when people of different age groups work together? It is often said that we learn more from our mistakes than our successes. Tell about a mistake that you once made and learned something from.
6/19/2010	5	Some parents believe that it is important for young children to focus on education and studying. Other parents think it is better that young children have lots of time for playing. What do you think is the best approach for the healthy development of young children? Give specific reasons and examples to support your opinion. In everyone's life, some days are particularly memorable. Describe what happened on one such day that you have experienced in your own life. What made it memorable and why? Include specific details such as where you were, who else was there, and so on, that will convince the reader how memorable the day was.

These pilot interviews were exploratory in nature. Few scripted questions were used so as not to lead the interviewees in any particular direction (DiPardo, 1994: 170). The writing booklet (see sample in Appendix 6), which included both the printed prompt and the written response that the interviewee had completed were used as the stimulus for the recall of the writing test. The only predetermined questions used were:

- Which prompt did you choose to write on in today's test?
- How did you choose which prompt to write on?
- What did you do after you chose the prompt you wanted to write on?
- How did you begin your essay?
- Did you use the writing prompt while you were working on your essay?

The interviews then unfolded in a way that was driven by the answers given by the interviewees. For example, responses from the interviewees that touched on prompt use or composing processes would be probed for more detail, with the aim of learning more about how the prompt was utilized during the writing test, without leading the interviewees in a predetermined direction. The researcher encouraged the interviewees to recall their thoughts and decisions from the writing test as "they were thinking at the time, and not what they thought they should have thought or done, or how they thought they should have responded," (Barkaoui, Brooks, Swain, and Lapkin; 2013).

The written response produced by the test taker during the writing test served as the basis for the interviews. For example, when test takers were asked about which prompt they had selected, they were shown the two prompts that were presented in the actual test booklet. When asked how they began composing their responses, they were shown the start of their own responses. Interviewees were given time to think and describe how they tackled the writing test with the interviewer (the researcher) prompting only when the interviewee had clearly finished her or his turn. The aim of these initial interviews was to elicit as full an account as possible from the test takers of how they engaged with the prompt and used it throughout the writing test.

There were 10 interviews in this initial exploratory phase of the work. The recordings were transcribed by the researcher and the transcriptions were read repeatedly in order to become familiar with the content of the interviews. Some themes that became apparent from these interviews were:

- The decision about which prompt to respond to was a serious one for the test takers
- Familiar topics were favored by test takers.
- Some form of essay planning was undertaken by most test takers.
- The time constraint was a serious concern.
- The layout of the test booklet and the ability to see the prompt throughout the test was important to the test takers.
- The test takers wanted to refer back to the prompt during the test.
- Some test takers were very focused on specific wording in the prompt.

7.2.3 Reflections on pilot of stimulated recall interviews

After reflecting on the points raised in the initial round of stimulated recall interviews, a more specific list of interview questions was formulated. As described by DiPardo (1994), the move from an initial open-ended round of stimulated recall interviews to a more focused follow-up phase can aid in eliciting responses from subjects that are more relevant to the research questions than is possible with a single round of data collection. While the initial round of stimulated recall interviews was exploratory, the interviews in Phase 2 explored in greater detail the themes that emerged from the initial interviews. New interview questions were created to focus on the issues that had emerged from the first round of interviews. The questions that were asked during the second round of test taker interviews were as follows.

1. Which prompt did you decide to write on?
2. What were your reasons for choosing that prompt?
3. Was there any reason you did not choose the other prompt?
4. Were there any words in the prompt that you thought were important?
5. What did you do next after you decided which prompt to write on?
6. How did you begin your essay?
7. Did you look back to the prompt while you were writing your essay?
8. How did you end your essay?
9. Did you feel that you had enough time to write your essay?

10. If you took the test again, would you choose the same prompt?

Table 7.3 shows how these questions related to the themes that had emerged from the interviews performed in Phase 1.

Table 7.3: Themes and interview foci that emerged from Phase 1

Question	**Theme from initial interview**	**Area of focus**
1	Importance of prompt selection/ familiar topics favored	Prompt wording
2	Importance of prompt selection/ familiar topics favored	Prompt wording
3	Importance of prompt selection/ familiar topics favored	Prompt wording
4	Focus on specific wording in prompt	Prompt wording
5	Some form of planning was undertaken/time constraint	Prompt use/composing processes
6	Time constraint	Composing processes
7	Test takers wanted to refer to prompt	Prompt use
8	Time constraint	Composing processes
9	Time constraint	All
10	Importance of prompt selection	Prompt wording/prompt use

7.2.4 Second round (phase 2) of stimulated recall interviews

The next rounds of interviews (phase 2) were conducted on four separate visits to the Toronto MELAB test center between September 2010 and October 2012. Table 7.4 shows the prompts that appeared on the MELAB during the administrations at which the interviews were conducted.

Table 7.4: Prompts used in Phase 2 of stimulated recall interviews

Date	Number of interviewees	Prompts
09/18/2010	4	In some countries such as the United States, the standard work week is 5 days, 8 hours a day. Some companies are considering changing their work week to 4 days, 10 hours a day. What are the advantages and disadvantages of the two options? Which do you think would be better for society? Give reasons to support your opinion. It is often said that we learn more from our mistakes than our successes. Tell about a mistake that you once made and learned something from
10/22/2011	5	Many North American professors complain that some students are not paying enough attention in class. Some suggest that the students are distracted and don't always study hard enough. What steps, if any, should be taken to address this problem? Explain, giving details to support your opinion. In everyone's life, some days are particularly memorable. Describe what happened on one such day that you have experienced in your own life. What made it memorable and why? Include specific details such as where you were, who else was there, and so on, that will convince the reader how memorable the day was.
3/24/2012	3	In some countries such as the United States, the standard work week is 5 days, 8 hours a day. Some companies are considering changing their work week to 4 days, 10 hours a day. What are the advantages and disadvantages of the two options? Which do you think would be better for society? Give reasons to support your opinion. Some child psychologists believe that the peer groups children play with influence their character and personality development more than the children's parents do. The psychologists say children are more interested in fitting in with their friends than behaving the way their parents want them to. Do you agree or disagree with these psychologists? Explain your point of view.
10/19/2012	6	As society changes, some professions become more important and others become less important. What job or profession do you think will be most needed in your country ten years from now? Give specific reasons and examples. Active vacations, for example where people go hiking or take a cooking course overseas are popular with many people. Others prefer to just relax and go to the beach. What do you think is the best way to spend your vacation time? Support your answer with examples.

All of the interviews were performed at the same test center because the Toronto MELAB test center had the largest and most diverse MELAB test population in North America. The test population in Toronto was representative of the MELAB population with test takers across the full proficiency spectrum of the test. Table 7.5 below summarizes the sample population of interviewees over the course of the four rounds of interviews.

Table 7.5: Sample population who participated in the stimulated recall interviews

Number of interviewees	Number of L1s	Average age	Total length of recorded interviews	Average interview length
18	12	26	201 minutes	11.17 minutes

All interviews were digitally recorded and the recordings were transcribed as soon as possible following the interviews The transcriptions were checked for accuracy by both the researcher and two colleagues who are familiar with the MELAB Writing Test. Following Brown (2007) and Green (1997) the transcripts were read in order to identify "a single or several utterances with a single aspect of the event as the focus (Brown, 2007: 111). The transcribed interviews were read repeatedly by the researcher to build familiarity with the content of the interviews. The process of familiarization with the content of the interviews allowed for some key themes to emerge. The coding of themes was an iterative process, reading for common utterances and concepts from the transcripts until some key themes became apparent (Brown, 2007). Due to the relatively small number of interviews and the fact that each interview was quite short (see Table 7.5 above), the transcripts were coded by hand. No software program was used in the analyses of the transcripts. The following themes were common to most interviews.

- Topic familiarity – comments were tagged as *topic familiarity* when test takers frequently cited the familiarity (or lack of) of the writing prompt topic as a principal reason why a particular prompt was chosen to respond to.

- Prompt wording – all comments made by interviewees that specifically referred to the specific wording within the writing prompt were tagged as *prompt wording*.

- Time constraints – comments were tagged as *time constraints* when interviewees referred to the influence of the time limit of the writing test.

- Planning – comments from interviewees about how they prepared to respond to the prompt, before they had begun to compose were tagged as *planning*. If the interviewee's writing booklet was marked up in any way (underlining of prompt wording or text from the writer that appeared to be planning), the interviewee was asked to explain the purpose of the inscriptions.

- Prompt use – when the interviewees talked about how they made use of the writing prompt and wanted to refer back to it throughout the composing process, these mentions were tagged as *prompt use*.

In addition to these main themes that emerged from the reading of the interview transcripts, further work with the transcripts while coding the categories above revealed further important themes. The themes that emerged from this second round of coding were:

- Composing – test takers described a range of strategies that they employed as they composed their responses. These strategies encompassed different techniques such as how the responses were initiated and concluded and how ideas were linked together. These mentions of strategies used while writing were tagged as *composing*.

- Structure – when interviewees made specific reference about how they organized their responses, they were tagged as *structure*.

- Vocabulary – test takers cited their lexical resources as an important factor in their decision making processes during the writing test. They gauged whether they knew enough vocabulary to choose a particular prompt and whether they had sufficient lexical resources to include certain topics in their responses. These mentions of lexical resources were tagged as *vocabulary*.

- Cultural difference – another feature that some interviewees raised was that of cultural awareness. Some prompt topics were said to be unfamiliar and these points were tagged as *cultural difference*.

The coding of the transcripts was completed by the researcher and checked by two colleagues who are familiar with the MELAB Writing Test. Following the review of the coding, the researcher met with the two reviewers. There was a discussion regarding how some of the comments had been categorized, with a particular focus on the categories of *time constraints*, *composing*, and *structure*. The reviewers questioned the coding of five interviewee comments that touched on how the time constraint had affected their ability to plan or refer to the prompt. After discussion, the coding was standardized so as to reflect the consensus understanding of when a reference to *time constraints* was tagged as such. For example, in the quote below from an interviewee, the comment was tagged as *time constraint* rather than *planning*.

> *I did have enough time to write but I mean I could've . . . It was short. It was short. But I had enough time because I sort of knew what to expect by how many pages they gave us to write so you kind of factor that in when you're making your outline. But, no I didn't think it was that much time to be honest.*

The interviewee talked about the interaction of the time constraint with the ability to plan a response. However, the interviewee's point was mainly focused on the effect of the time constraint with planning as a supporting issue. Hence, in this case there was sufficient focus on the aspect of the time constraint

for it to be tagged as such. After the review of coding, as described above the researcher and two reviewers reached consensus agreement on all categories and coding.

The total number of times each of the themes was mentioned was tallied. The results of this coding are presented in the following section.

7.3 Results

Table 7.6 shows the number of mentions for each of the nine major points raised by the 18 interviewees during the stimulated recall (SR) interviews. The results show that time was the point most frequently referred to by interviewees. Topic familiarity, and prompt wording were the two other points that were mentioned 30 times or more. Cultural difference was the point that was mentioned least frequently.

Table 7.6: Number of mentions of points raised during SR interviews

	Number of mentions	Number of interviewees
Time constraints	34	15
Topic familiarity	32	16
Prompt wording	30	10
Composing	28	16
Prompt use	20	15
Planning	20	11
Vocabulary	17	7
Structure	16	12
Cultural difference	7	5

The following sections will address the comments made by the interviewees and examine the themes common to the interviewees' comments on each point.

7.3.1 Time constraints

The test takers who participated in the SR interviews made multiple references to the time pressure they felt they were under during the MELAB Writing section and how the time constraint impacted their test taking processes. 15 of the 18 interviewees made reference to time constraints. Five interviewees suggested that the 30 minutes allowed for the writing section was insufficient.

03241201; L1 Malayalam

> *I don't have time to like, you know, I don't have time to tell a lot of beautiful something like I write another essay because in 30 minutes I have to go with the main point and explain it.*
>
> *For example, I wrote 3 statement at the beginning. I couldn't even finish my 3 statement within the time. I finish only 2 but the last one I won't be able to do the last statement so I have to conclusion and my conclusion, I couldn't have like, I know the end of the essay might important because it's like the impression at the end but you know it's impossible because I don't even have time.*

10191204; L1 Punjabi

> *Yeah, so like 30 minutes is like, you have so much pressure. Even if you know the things you want to write, you just keep looking at time.*

Although there were multiple comments (34 in total) about the impact of the time limit, 24 of these comments made reference to how the time limit placed restrictions on choices the interviewees would like to make and test taking strategies they would like to employ but felt unable to because of a shortage of time. For example, interviewees commented on how the time limit influenced the prompt they chose to write on and how they planned their response.

10221101; L1 Yoruba

> *Yeah, because of the time constraint I prefer this and I wrote on it.*
>
> *You know, in the normal conditions you can just, you know outline and from the outline you can place them in order but in an exam you don't have that time, you just have to write.*

10221102; L1 Mandarin

> *I'm just having hard time thinking of all this things and putting the answer to the question, like putting them all together it feels like 30 minutes is not enough to think of all those.*

There were seven comments indicating that even though the MELAB offers test takers a choice of prompts in the writing section, interviewees felt that some prompts could not be selected because a response could not be produced in 30 minutes. The following section (Section 7.3.2) will show how important the familiarity of the topic presented in the prompts was when selecting a prompt. There is clearly some interaction between the time limit the test takers are working under and the familiarity of the topics that test takers are presented with.

Four interviewees stated that they would have liked to do more planning before beginning to write but they felt that was not possible because of the time limit. Seven interviewees also commented on how the time limit influenced their composing process and how they used the prompt during the composing process.

03241203; L1 Farsi

> *I don't go back to the prompt but actually I have second idea that I need to develop here and I don't have time to develop it but I want to mention it with the other subjects.*

10221105; L1 Ukrainian

> *I think I can use even like some better sentence to go from this paragraph to this paragraph but it needs more time.*

As will be discussed in greater depth in the section on prompt use (7.3.5), interviewees often described how they utilized the writing prompt during the composing process. This generally took the form of

looking back to the prompt while composing the response to check that the response was on target. Five interviewees who said they used the prompt in this way added that they would have liked to look back to the prompt more frequently as they wrote but felt this was not possible due to time constraints.

As shown in the excerpts above, some interviewees felt that they were not able to fully demonstrate their writing proficiency because of the time constraint. Interviewees felt that they may be able to write on other topics, use more sophisticated vocabulary, connect ideas together differently, and write more complex sentences if they were allowed more time. While there is no direct evidence that test takers would perform better (in terms of being awarded a higher score) given more time, there were comments from nine interviewees that the 30-minute time limit restricted their performance on the writing section. These comments, heard consistently from multiple interviewees indicate there is a possibility that the time constraint negatively affects performance on the writing section.

In terms of writing prompt characteristics, as will be apparent from the following section (Section 7.3.2), prompt topics that were unfamiliar to the test takers were likely to be rejected, at least partially because it was challenging to compose a response within the time constraint. Also, as will be discussed in Section 7.3.3, writing prompts that had a large number of rhetorical cues were seen as easier to organize and plan a response to because the rhetorical cues provided the frame for the response. As the time constraint affected the interviewees' ability to plan, it seems likely that the time constraint encouraged test takers to select prompts with several rhetorical cues over ones with a smaller number of cues.

7.3.2 Topic familiarity

During the stimulated recall interviews, all but two of the 18 test takers interviewed commented on the importance of topic familiarity and 32 mentions were made of this point in total. Comments were tagged as *topic familiarity* when they specifically referenced the interviewees' level of familiarity with topics or subject matter presented in the prompt. The interviewees were very consistent in saying that familiarity with the topic was the most important point to consider when selecting a prompt. Several of the interviewees emphasized the issue of topic familiarity, as shown in the quotes below.

10221102; L1 Mandarin

> *This kind of question is err (pause) too beyond my daily life, you know. But the second one is more personal and I can write with my feeling.*

10221103; L1 Tagalog

> *Because, for me it's much easier for me to express myself if it will relate on my experience. The letter A is . . . I'm gonna' have more of a hard time doing the, setting the thesis and like that. But because I have like this personal experience, it is like I will really have chance to, hmmm, for me to pass the essay or have a higher rating because I really know the topic, yeah and that's why I decided to do this one.*

During the interviews, there was a sense that the choice of which prompt to respond to was a very important one. Several interviewees went to great length to explain why they had selected one topic and not the other. Interviewees who had the opportunity to write on a topic they felt very familiar with, expressed how fortunate they thought themselves. For example, interviewees who were medical or education professionals (taking the MELAB for licensure purposes) spoke of being able to respond to a prompt that focused on health or educational topics. The interviewees felt that they could fully utilize their content knowledge and vocabulary to write a lengthy response to the prompt.

10191206; L1 Arabic

> *Yeah, I mean, absolutely, like if there was a question about anything that's to do with the medical field I think I would ace that. I mean it really depends on what you experienced and what you went through. I mean if you ask me a question about like history, I mean I don't know anything. So, I mean yeah, certain questions you can really elaborate and you can really bring out everything into the essay.*

10191204; L1 Gujarati

> *I can just tell you about myself because being a teacher, I've been teaching for nineteen years. Then I thought choosing topic A will help me to highlight my educational practices more than topic B. So, then I thought I'll get more points, more information regarding that job, that is my profession to have like my teaching profession*

A lack of familiarity with the topic presented in certain prompts interacted with time constraints for some interviewees. They felt nervous about tackling certain topics because they felt that they would need to spend too much time thinking of ideas for content in their responses. The lack of time available to them in the test discouraged them from choosing such topics as they were concerned that they would not be able to produce a sufficiently long response within the allotted time.

10221101; L1 Yoruba

> *And the second one is, you know, based on the memory. I have a lot of memories but I can just pick on one and just write, irrespective of the time. That isn't enough. But you have to make important points. It's not a preamble and you know, at the end of the day you've got to know where you're going. You don't have much time so that's one of the reasons.*

Another point that came up repeatedly was a preference for personal domain topics, which allowed interviewees to write on topics that they were very familiar with.

10221104; L1 Gujarati

> *It's about my life, so, as everyone have, I have many memories for my life and I can describe that easily so it gives me point to write down, what to write and how to paragraph the essay and manage the essay. I have my own control on that essay.*

10221103; L1 Tagalog

> *Sometimes they said that you can just make up stories but it's still easier if it's based for your experience.*

Six interviewees stated a preference for personal domain topics because they felt confident that they would be able to write a long response. Five interviewees stated that they felt that the length of their response was important and that they had a better chance of getting a high score if they could write a lot. Topics in the personal domain were associated with the opportunity to write a lengthy response on a familiar topic and get a good score on the writing section and this was a view expressed consistently throughout the stimulated recall interviews. The interviewees made it apparent that topics in the personal domain were viewed as easier than topics situated in other domains.

Overall, the importance of the topics presented in the choice of prompts was evident from the stimulated recall interviews. Prompts in the personal domain were preferred by many test takers and viewed as easier than other prompts. Prompts that covered topics test takers had educational or occupational experience of were also favored over prompts that were unfamiliar. Topics that were unfamiliar to test takers were not selected, in part because to the time constraint. Interviewees made clear that they feared they would not be able to think of enough ideas to write a response they felt was sufficiently developed.

7.3.3 Prompt wording

Prompt wording was the third most mentioned point in the interviews overall with 30 mentions. This category refers to comments made by interviewees specifically about words used within the writing prompt presented on the test booklet. Ten interviewees reported using the prompt wording in both the planning and composing stages. In all the stimulated recall interviews, the interviewees' writing section answer booklet was used to stimulate the individual's recollection of the test itself. It was apparent that all but four of the interviewees (24 in total) had underlined or similarly marked up the writing prompt with attention given to certain words.

The interviewees reported specific words or phrases within the prompt that they paid particular attention to when planning or composing their responses.

10191201; L1 Punjabi

> *Hmm, I wrote, thought the word, hmmm, "important" was important. "About 10 years" was important so that it put the word future in perspective. Hmm, did they use the word future? No. But you know, to have an idea about time.*

10191204; L1 Punjabi

> *I did, "most needed in your country." This helped me a lot to deliver my essay in such a way so that, because in India there are, our country is 90% literate. If we implement new strategy it may help for the some percent illiterate people. So, that helped me.*

Comments such as these from the interviewees indicate that they pay close attention to the specific wording used in the prompt. They often report giving careful thought to the words in the prompt that they must consider as they planned and composed a response. The interviewees' interest in the specific wording within the prompt seemed to be motivated mainly by a wish to compose a response that was as relevant as possible to the task set in the prompt. The volume of comments (30 in total) from all interviewees on this point indicates the importance of the specifics of prompt wording.

Although the number of words that were marked up varied greatly, with some interviewees marking only three or four words and others marking many more, the words that were most consistently marked were content words and rhetorical cues. The interviewees were not marking language they found hard to process in the prompt. At no point did interviewees talk about any difficulties with understanding the grammar or vocabulary in the writing prompts. Rather, the marked words were the ones that interviewees considered to be important in terms of where the focus of the response should be targeted via specific content words or task wording (for example; *give your opinion*, or *discuss the advantages and disadvantage*s). These are the rhetorical cues within the prompt that are one of the core characteristics established in Chapter 4.

A specific comment made about prompt wording by five interviewees was that having a large number of rhetorical cues was helpful. The rhetorical cues acted as a guide for how the response would be organized and this guidance allowed the interviewees to save time when planning and composing.

10221103; L1 Tagalog

> *They have lots of questions it help me build up my essay even better because if it will just, if it is just "what made it memorable and why" and didn't do other questions, I might not be able to put 250 words together, you know. So, for me, like hmmm, answering more questions is much better in creating an essay.*
>
> *Yeah, and you can really express what you really want with all those questions instead of just focusing on one.*

03241202; L1 Arabic

> *I like the full explanation. Even though I was choosing this topic everything was there if I was breaking down it, everything is here. I can make like 5 or 6 paragraphs.*

These interviewees stated a preference for a writing prompt with a large number of rhetorical cues. None of the interviewees described prompts with more cues as being more challenging or complex. The interviewees who stated a preference for such prompts wanted to be presented with more cues as they helped guide and support a lengthier response.

In summary, the interviewees described the importance of the wording within the writing prompt. The interviewees did not report any difficulties understanding the wording within the prompts; there was no mention of words or phrases that caused difficulty for the interviewees. Rather, the interviewees saw the importance of the rhetorical cues within the prompt and specific wording that guided them in

planning and composing their responses. The interviewees used the prompt wording strategically, in an attempt to decode the intentions behind the prompts and guide them in creating a response that would maximize their opportunity to achieve a good score.

7.3.4 Composing

16 interviewees made reference to their use of composing processes during the stimulated recall interviews. In total, composing was mentioned 28 times by the test takers. Comments about composing referred to strategies test takers adopted or decisions they made as they were composing their responses, such as how they initiated a response or how they made certain decisions while responding to the prompt.

10191202; L1 Tagalog

> *I just sort of started going with it. It was something recent for me so it was really easy for me to know the three things that I enjoyed so I just jumped into it. Oh, and I explained that there's a variety of different types of vacations like it is explained here so I took the question here and explained it in here and expanded on all of it. So essentially just took the question, pointed out three factors in that question and wrote about it.*

10191206; L1 Arabic

> *Not just write a lot of content but also like you know, if there's an argument you can say like you know that this is good because of blah blah and, however, like you know because it's not good because this and that. But I can also give my opinion so here's only opinion based but there you're actually describing two different points and then provide your opinion*

Typically, the strategies that interviewees spoke of were based on their own personal beliefs regarding the type of response they needed to produce to get a high score on the writing section. These beliefs cover various writing strategies, such as how the response should begin, the lexical choices that should be made, and how the conclusion should be handled. There was little consistency in the interviewees' comments on composing; they chose different aspects of the responses to refer to and the strategies they adopted seemed to be personal preferences.

The Toronto test center where these interviews were conducted, offers a preparatory workshop dedicated to the MELAB Writing section but there was no evidence from these interviews that the interviewees were using comparable composing strategies, learned at the workshop. The interviewees did not consistently talk about typically taught strategies such as presenting a clear thesis statement, developing a five-paragraph response, or using topic sentences. The only aspect of the approach to composing that was common among the interviewees, was that they were employing strategies they felt would make a positive impression on raters and would hence help them achieve a better score on the writing section. These comments from the interviews reveal that while these individual test takers do indeed consider how they compose their responses, they are apparently not formulaic in how they

approach the writing section. That is, systematic test taking strategies do not seem to have been adopted by this group of interviewees.

7.3.5 Prompt use

All but three of the 18 interviewees commented on how they used the writing prompt during the writing test and 20 comments were recorded in total. The majority of these comments referred to looking back to the prompt during the planning and/or composing processes and how the interviewees utilized the writing prompt itself during the act of taking the test. Several test takers commented on using the prompt throughout the test, making reference to it regularly to guide both their plan and their response.

10191201; L1 Punjabi

> *I noticed near the end that I didn't, the prompt itself said, it specified like talk about your country and so in the end I just threw in the word Canada just to make it very clear that I understood what the question was.*

10191205; L1 Spanish

> *Yes, when I'm stuck. I would just go back to search for things, go back to my brainstorming and that's it. To write next, how to make it longer, how to . . .*

Although the test takers had different purposes for referring to the prompt as they planned or composed, almost all the interviewees described using the prompt to some extent throughout the test. Different test takers used the prompt at different times; almost all of them used the prompt during the planning phase, many referred back to the prompt while they were composing and others reported using the prompt toward the end of the writing test to check that their response had properly addressed the prompt. Regardless of when test takers opted to refer to the prompt, it was apparent that it played a key role throughout the test taking process. Being able to easily access and refer to the prompt throughout the test administration was important for test takers.

The physical and/or virtual design of the test (depending on whether the test is delivered in paper and pencil format or is computer/internet delivered) needs to take into account test takers' preferences for regularly referring back to the prompt at any given stage of the test taking process. As with the comments reported above on prompt wording, these comments on prompt use show that the interviewees were focused on the prompt throughout the writing test. The need for clear and unambiguous prompt wording and a test design that allows test takers to view the prompt whenever they wish to are both important for test takers, based on the accounts provided in these interviews.

7.3.6 Planning

11 of the 18 test takers interviewed talked about planning their response and there were a total of 20 mentions of planning. The types of planning strategies that test takers reported using included writing an outline prior to composing, underlining key words in the prompt, and thinking carefully about the essay outline before beginning to write.

10221102; L1 Mandarin

> *While I'm writing, at the beginning of the writing I keep looking at these important point so I make sure I include them in my passage. But after that this is from 5 minutes you see. After 5 minutes I have nothing on here.*

10191204; L1 Punjabi

> *I was thinking of how to convey my message. Like, what should I do in introduction, for the body and the conclusion? I had, you know a vague information. I was thinking of something and then I started.*

There were few clearly discernible trends in the comments related to planning. The desire to plan was common but the way the planning was undertaken differed widely. What was of most interest was that planning was undertaken by these test takers at all. The interviewees felt a great deal of time pressure, as is clear from Section 7.1 above. However, despite feeling that 30 minutes was insufficient to complete the writing task, a majority of the test takers still attempted to plan before beginning to compose. The planning was minimal and often consisted of some simple underlining of words in the prompt or a quickly sketched outline, but these interviewees still saw planning as important.

The fact that a majority of these interviewees stated a desire to plan before they begin composing is some indication that planning time may be a positive addition to writing tests. These interviewees repeatedly talked about the time pressure they felt they were under. Adding a few extra minutes to the writing section, when only planning is allowed would give test takers the opportunity to make a more considered decision on which prompt to respond to. The additional time for planning would also give test takers what most of these interviewees say they need; more time to think and prepare their responses before they begin to compose.

7.3.7 Vocabulary

Seven of the 18 test takers interviewed commented on the importance of vocabulary and there were a total of 17 mentions of vocabulary. Mainly, these comments made reference to the lexical resources that the interviewees felt they needed in order to be able to fully respond to a particular prompt. The comments were not directed at the vocabulary used in the prompt itself (the interviewees almost never reported finding the prompts difficult to understand) but rather, at the words they feel they need to know (but often don't) in order to be able to write on a particular topic. That is, comments were tagged as vocabulary when they were related to the words interviewees felt they needed to know to respond to a particular prompt.

10221105; L1 Ukrainian

> *I found that here on this topic, I cannot use more words which I have in my vocabulary. So, that's why I was not satisfying with this two topics very much because I cannot express my language very much.*

03241202; L1 Arabic

> *Because parents might think differently than they think before so that's what I applied to write this essay and first of all what I did is like I make some list for my own vocabulary because I don't want to, that was the only thing why I chose this topic so I listed my vocabulary.*

As the interviewees wanted to write on a topic they were familiar with, they also wanted to write on a topic they felt they had sufficient lexical resources to express themselves on. Interviewees seemed to rule out certain topics due to a lack of vocabulary. Any topic that interviewees felt they would not be able to express themselves on was likely to be rejected. The comments made by these interviewees on topic familiarity and vocabulary provide some clear evidence regarding the prompt characteristics that are considered when selecting a prompt. Ideally, the interviewees wanted to select a prompt that is based on a topic that is familiar to them in terms of content. They also wanted to avoid responding to prompts that are based on topics that require specific vocabulary that the test takers lack. Hence, the topics that are presented in writing prompts and the vocabulary that is needed to respond to them must be carefully considered by the writing prompt developers. The pretesting of these prompts on a representative pilot population is also necessary as the vocabulary that is needed to respond to a prompt will not always be evident. Prompts that advantage or disadvantage certain test takers due to the need for context-specific, specialized or technical vocabulary should be identified at the piloting phase.

7.3.8 Structure

12 of the 18 interviewees commented on structure and this point was mentioned 16 times in total during the interviews. Comments were categorized as referring to structure when they focused on identifiable parts of the response such as how the writer handled the introduction, or the conclusion, or linked together paragraphs. This point is distinct from planning and composing in that it deals specifically with the organizing pattern of the response.

03241202; L1 Arabic

> *So, I just give like an example of what children are and what they do and then I just focused on like related to my main topic like what they say and how I agree or disagree because I wanted to say in the introduction that I agree with this topic. That is not meaningful if I'm writing it in like second or third paragraph now it is ahead of the whole story.*
>
> *So when you finish your introduction you have to support it and supporting it you need to know now what you are supporting.*

10221103; L1 Tagalog

> *I just look at the things I want to discuss in my first paragraph and then I just develop it and enhance it more in my second, third, and fourth paragraph and so on and make it more clear to the person who's going to read it and convince him or her that it was a good essay.*

Many of the interviewees talked about how they think, sometimes quite carefully, about the structure of their responses. However, they reported that their efforts to produce a well-organized response were sometimes compromised by the limited time that they had available during the writing section. What is clear from the interviews is that the interviewees cared about the organization of their responses, just as they tried to plan the responses before beginning to compose. Although the interviewees were consistent in their comments about the restrictions of the time limit, they still made efforts to plan and organize their responses.

7.3.9 Cultural difference

The final point made by the interviewees was related to topics that they found culturally unfamiliar. Comments were tagged as *cultural difference* when they referred to difficulties interviewees had understanding certain prompt topics because they were culturally unfamiliar. Only five test takers made such comments for a total of seven comments but those individuals who did talk about cultural differences often did so in the context of expressing their feeling that some test topics seemed unfair and inaccessible to them.

10221104; L1 Gujarati

> *Because it is about the education system in America, like North America and here it is different than our country so I'm not used to that. Like in India, I'm from India, so we are like very obedient to the teacher, we cannot do anything. We cannot talk, we cannot use mobiles, we cannot do anything, so I don't have much idea for this to describe about how to discipline the students.*

10221103; L1 Tagalog

> *I am from the Philippines so for me, basically I don't really have idea about the North American professors and how they handle students.*

These comments were most commonly made about topics situated in the educational domain. The MELAB is commonly taken for higher-education admissions purposes for those who want to study at an English-language medium university in North America. MELAB test takers tend to lack familiarity with the norms of university life in the USA and Canada. Writing prompts that require test takers to discuss or argue for a particular educational approach in the higher-education context can be frustrating for these test takers. Topics situated in the educational domain that focus on aspects of higher education have now been added to the list of topics that are flagged and revised during the fairness and bias review stage of development for the MELAB.

The findings related to *cultural difference* are consistent with those reported about *topic familiarity* (Section 7.3.2) in the sense that test takers prioritize topic familiarity when they select a prompt and reject topics that are unfamiliar to them. Familiarity is a strong influence on prompt choice.

7.4 Summary

The stimulated recall interviews with the test takers revealed that certain prompt characteristics had an effect on the prompts that interviewees selected, and on the interviewees' planning and composing processes. The interviewees emphasized that prompts with familiar topics were favored over other prompts. The interviewees wanted to respond to prompts based on a topic that they felt they were able to write at length on. They felt that these prompts gave them the best opportunity of writing a long response and performing well on the test, in terms of the score they may be awarded. The interviewees also wanted to respond to a topic that allowed them to demonstrate the full range of their lexical resources. Topics that the interviewees felt unfamiliar with or did not have sufficient vocabulary to develop a response to were rejected. Prompts situated in the personal domain were specifically referenced as being easier to respond to than prompts situated in other domains. These prompts allowed the interviewees to write on a personally familiar topic and to fully utilize their lexical resources. The other prompt characteristic that was favored by test takers was prompts with a large number of rhetorical cues. Prompts with a large number of rhetorical cues were seen as easier than other prompts because the cues provide the structure for the response without the writers having to organize their own ideas.

The issue of time constraint was one that was emphasized by many interviewees. A majority of those interviewed said that they felt they did not have enough time to complete the writing section. The lack of time influenced the prompt they opted to respond to, how they planned their response, the ability to use the prompt throughout the writing test, and the composing strategies employed during the test. Although the time limit is the same for all test takers, it is clear that many interviewees are uncomfortable with the restrictions they believe it places on their capacity to demonstrate their writing proficiency. The time constraint influenced, to some extent, most of the other factors that the interviewees talked about during the interviews.

Despite concerns over the time limit, the interviewees reported their efforts to plan a response, to effectively structure a response, and to utilize the prompt throughout the test in order to keep the response on target. The time constraint does not stop them from utilizing certain writing strategies (planning, structuring, using the prompt to stay on target). However, many interviewees would have liked to make more use of these strategies but could not do so because of the time constraint. The interviewees also reported focusing on certain key words within the prompt while they planned and composed their responses. This is an indication of the importance of prompt wording, given that some interviewees reported latching on to certain words in the prompt and using them at a variety of stages during the test.

Finally, a small number of interviewees said that they found some prompts to be culturally unfamiliar and inaccessible. These prompts were avoided as the interviewees felt that they lacked sufficient knowledge of the topic to write a response they could have confidence in. The interviewees described prompts in the educational domain as sometimes being unfamiliar to them, especially those that asked about aspects of higher education. This finding can perhaps be attributed to specifics of the test

population but it is an indication that test programs and test populations may have 'blind spots' or particular topics that are unsuited to being used on the test.

7.5 Implications

There are implications from the findings of these interviews for the design, development, and administration of writing tests, especially those with high-stakes consequences. From the aspect of test design, these findings emphasize the need for a clear construct definition of second language writing proficiency. Without a clear definition of what aspects of writing proficiency are to be assessed, it will be very challenging to make defensible decisions on which topics are permissible and appropriate for use on the test. For example, if the test construct is defined as a measure of academic writing proficiency, then topics that are not situated in an appropriate domain can be excluded from the test specifications. Similarly, the more precisely the construct can be defined, the easier it is to design prompts that operationalize the intended construct of writing proficiency. Without this clear definition of construct, the topics and tasks that are not appropriate for the test are more challenging to identify.

The interview findings reveal that certain topics, especially those in the personal domain are seen as easier than topics in other domains. In addition, prompts with a large number of rhetorical cues are also seen as easier than prompts with fewer cues. These findings indicate that test designers may wish to prohibit the use of personal domain topics and prompts with a large number of rhetorical cues as these prompts may contribute to problems with prompt equivalence. Removing prompts that are seen as easier than others by the test population will help make the pool of operational prompts more consistent in terms of their equivalency. Crafting prompt specifications that prohibit the development of personal domain prompts and prompts with large numbers of rhetorical cues will make a positive contribution to maintaining prompt equivalence.

The interview findings also clearly show the impact of the time limit on the interviewees. While there is no evidence that the time limit will impact the score awarded, the comments made by these test takers suggest that scores may be impacted based on the time allowed for the writing section. Seat time and issues of test taker fatigue are very real concerns and practical constraints are likely to mean that high-stakes tests with large candidatures need to enforce a time limit. However, that time limit must be set appropriately and be based on empirical evidence that the time restriction does not prevent test takers from optimally demonstrating their writing proficiency. Test designers may also wish to consider including a set period of planning time as part of the administration procedures.

The high level of engagement that test takers say they have with the wording of the prompt reinforces the need for a thorough test development process with multiple levels of review. The wording of the prompt needs to be accessible and transparent to all test takers. Prompts that have been through a comprehensive review process, as is the case with the MELAB, are more likely to be able to avoid problems of vague or confusing prompt wording. Prompts should also be pretested on a representative pilot population before being used operationally. Pretesting will allow topics that are inaccessible, inappropriate, or uninteresting for the test population to be identified and removed from the pool of

operational prompts. Regular pretesting allows test developers to build up an understanding of which topics are unsuitable for the test population and this understanding can be built into the prompt specifications and guidelines that are used by developers and reviewers.

Finally, the design of the physical or virtual layout of the test needs to take into account test takers' preference to be able to refer to the prompt throughout the writing test. A layout that hinders access to the prompt, at any point of the test will disadvantage the writers who want to look back to the prompt regularly.

The third research question presented in Chapter 4 asked:

- How do writing prompt characteristics affect the test takers' test taking processes?

The findings of the stimulated recall interviews indicate that prompt characteristics can indeed affect the prompt selection process, the planning process, and the composing process. Several different prompt characteristics can, for some test takers affect how they select a prompt and then plan and compose their responses. The familiarity of the topic (operationalized here by domain), the number of rhetorical cues, and the focus of the prompt are prompt characteristics that can have an effect on the test takers' test taking processes.

Chapter 8 – Discussion of findings from quantitative and qualitative studies

The findings from the quantitative approach were presented in Chapter 5 and discussed in Chapter 6. Similarly, the findings from the qualitative approach (the stimulated recall interviews) were presented and briefly discussed in Chapter 7. In this chapter the findings from the mixed methods approach (quantitative and qualitative strands) will be synthesized and compared with findings from previous studies in the writing assessment literature. The aim of this chapter is to situate the findings from the current study within the writing assessment field and to highlight findings that are original. Findings from similar studies will be compared with the findings from this research to clarify how the findings reported in this study differ from those in the writing assessment literature. After reading this chapter, there should be a clear understanding of the original contributions made by this study.

8.1 Findings common across the quantitative and qualitative approaches

The findings from the quantitative and qualitative approaches revealed some common prompt characteristics that have an effect on both the written product and the test taking process. The prompt characteristics that were common across both methods were prompts situated in the personal domain and prompts with an open focus (prompts that did not constrain the writer).

8.1.1 Prompts in the personal domain

The post-hoc analyses from the MANOVA showed that prompts 73 (mistakes) and 115 (memorable days) were responsible for a large proportion of the significant differences in written products (see Section 5.8.4). Both of these prompts are situated in the personal domain and these prompts elicited significantly different written products from prompts in other domains. MANOVA analyses showed that responses to prompts situated in the personal domain elicited responses that are lexically more sophisticated (see p.98-100) and syntactically simpler (see p.102-104) than responses to prompts in other domains. Responses to prompts in the personal domain also elicited lower levels of academic language use than prompts in other domains.

These findings from the quantitative analyses were supported by the findings from the interview data as interviewees also confirmed that prompts situated in the personal domain were viewed differently from prompts in other domains (see Section 7.3.2). Prompts that were situated in the personal domain were emphasized as being easier to respond to than other prompts, as test takers believed that they would be able to write a lengthy response, which would give them a better chance of getting a high score. The combination of these findings from both approaches marks the personal domain as a prompt characteristic that is likely to contribute to problems with establishing prompt equivalence. Hence, it is a prompt characteristic that likely needs to be controlled.

There have been similar findings reported in the writing assessment literature regarding concerns over personal domain prompts. Powers & Fowles (1988) learned that the interviewees consistently reported

that prompts that allowed test takers to draw on personal experiences were preferred and viewed as easier than other prompts. Topic familiarity was reported as the single most important factor when test takers selected a prompt, a finding that is comparable to the findings in this study.

Lim (2010) also described topic domain as the one prompt characteristic that yielded a significant difference in score awarded but the domains he reported as yielding significant differences were ones situated in the public and occupational domains. However, other studies have failed to find any effect of domain on written product. Brossell & Ash (1984) administered personal and neutral prompts to students in US colleges but found no significant difference in holistic scores awarded to the responses. Hoetker & Brossell (1989) found similar results for personal versus neutrally worded prompts. Few recent studies of prompt effect have investigated the effect of prompts in different domains on the written product. Instead, recent studies have tended to focus on prompts of differing levels of cognitive complexity. The findings reported in this thesis suggest that the topic of the prompt and the domain it is situated in can have a significant effect on the written product and this finding is echoed by the evidence presented by the interviewees who clearly stated the importance of topic familiarity when selecting a prompt.

8.1.2 Prompt focus

The results of the quantitative and qualitative phases of this work indicated that the original formulation of the open/focused distinction within the prompt categorization framework required some rethinking. The original distinction drawn between open and focused prompts was based on the degree of contextualization or support provided to the test taker by the prompt. Hence, prompt 115 (memorable days) was originally categorized as focused because of the specific cues provided to the test taker within the prompt that focused the content of the response. However, the analysis of the responses to this prompt (Section 6.3.1) revealed that prompt 115 actually allowed test takers to select from a very broad range of personal experiences when preparing a response. Prompts that elicited such a broad range of responses proved to be a key prompt characteristic that contributes to significantly different written products. The important distinction that needs to be captured in this prompt category is the extent to which test takers are free to respond on a broad range of possible response topics, versus whether they are constrained by the prompt to produce a far more restricted range of possible responses. This distinction is well covered by Lim (2009) in the writing assessment field and Leaper & Riazi (2014) in the speaking assessment field. Both approaches define the focus of the prompt by the range of possible topics that may be addressed while still providing an on topic response to the prompt wording.

The reading of the test taker responses to prompts 73 and 115 showed that they shared a key characteristic of not constraining the writer when composing a response (p.121-122). The broad range of topics elicited by both these prompts was vital to understanding the open/focused distinction. This reconsidered distinction in prompt characteristics was strongly supported by the stimulated recall interviews. Interviewees consistently reported that they favored prompts with familiar topics (Section 7.3.2). Prompts 73 and 115 allowed test takers to respond to a broad range of possible topics and this open focus, or unconstrained nature of the prompts makes them seem easier than focused or constrained prompts.

Open focus prompts elicited responses that were lexically richer (see p.115) and syntactically less complex (see p.135) than responses to other prompts. They also elicited responses with low levels of academic language use (Section 5.8.2). Conversely, focused prompts elicited responses that were syntactically more complex and lexically less sophisticated than responses to other prompts. The two open focus prompts that elicited responses with significantly different written products from the other four prompts were both situated in the personal domain. Tentatively then, prompt focus is another prompt characteristic that may be controlled for in order to improve prompt equivalence.

To the researcher's knowledge, the only other study that has explored a similar prompt distinction is that of Lim (2009: 97), who proposed a constrained/unconstrained distinction that is almost identical to the reconfigured view of the open/focused distinction in this research. However, Lim reported that the constrained/unconstrained distinction did not have a significant effect on the response, in terms of the scores awarded. Indeed, the results in Lim's study did not even approach significance for this prompt characteristic. As with the similar discrepancies between Lim's findings and these findings, the contrasting findings between the focus on test scores and the focus on textual features of the responses are quite stark. Lim (2009: 111) conducted ANOVAs on six prompt characteristics, with the prompt characteristic as the independent variables and the fair measure averages (from the Rasch model) as the dependent variables. The discourse analytic approach consistently shows significant differences in written products when the use of score as the criterion measure shows non-significance.

Both the quantitative and qualitative methods found two prompt characteristics (domain and focus) to have an identifiable effect on the written product. Prompts 73 (mistakes) and 115 (memorable day) are situated in the personal domain and have an open focus. These findings indicate that for writing prompts to be equivalent, prompts situated in the personal domain should not appear on the same test program as prompts situated in other domains. Similarly, prompts with an open focus should not be used on test programs alongside prompts that significantly constrain test takers in the topics they can write on. Test taker interviews and analyses of the written products indicated that these prompt characteristics elicit responses that differ from responses to prompts with other characteristics. The triangulation of these findings leads to the recommendation that these characteristics should be controlled for if prompt equivalence is to be established.

8.2 Findings from the quantitative approach

Section 8.1 presented findings that were common across both the quantitative and qualitative approaches. This section will address additional findings from the quantitative study. One other prompt characteristic, beyond domain and focus was associated with significantly different written products based on the MANOVA analyses. As shown in Section 6.4 the other prompt characteristic that elicited significant differences in the quantitative study, but was not highlighted in the qualitative approach was the response mode.

8.2.1 Response mode

The MANOVA analyses revealed that the narrative and argumentative responses differed significantly (see Table 5.25 on p.114) in a range of discourse measures, including those that operationalized lexical

sophistication (p.108), syntactic complexity (p.112-113), and academic language use (p.110-111). Narrative responses were significantly more lexically sophisticated than argumentative responses but narratives also elicited significantly lower levels of academic language use. Argumentative responses were significantly more syntactically complex than narrative responses.

A limited number of studies have investigated how different response modes affect written products in an assessment context. Quellmalz, Capell, & Chou (1982) found that narrative and expository responses differed significantly by the analytic scores awarded to the responses. Narrative responses were scored significantly lower than expository responses on three of the five scoring criteria, suggesting that narrative writing was harder for the participants than expository writing. Hamp-Lyons & Mathias (1994) also investigated the effect of response mode on holistic scores and concluded that the score awarded varied by the prompt characteristic of response mode.

Expository responses were not included in this research as the evidence provided by an analysis of responses by raters when determining the prompt categorization framework suggested that narrative and argumentative responses differed most distinctly and were the two response modes most worthy of investigation (see section 4.2.2.3). A study that brings in expository responses to compare with both narrative and argumentative responses would be an interesting direction for future research

There were no significant differences in scores reported in the current work for any of the prompt characteristics, including response mode, in contrast to the findings of Quellmalz, Capell, & Chou (1982) and Hamp-Lyons & Mathias (1995). Although the findings of these studies are not directly comparable to the findings of this thesis, they both indicate that response mode is a prompt characteristic that should be controlled for in order to establish prompt equivalence. The findings of this study indicate that prompts that elicit narrative responses and ones that elicit argumentative responses should probably not be used on the same test program.

Lim (2010) included response mode (or rhetorical task in his terminology) within his prompt categorization framework. He concluded that there were no significant differences in scores awarded to responses in the expository, argumentative, or narrative modes. Given that both Lim (2010) and the current study were performed in the same assessment context (the MELAB), the differences in methodology must be responsible for the inconsistent findings. Two major differences between Lim's work and the current study are:

1. Lim used the score awarded to the responses to interpret prompt effect while the current study takes a discourse analytic approach.
2. Lim analyzed 60 different independent writing prompts whereas the current study looks at only six prompts that differ markedly from each other.

These differences in approach may account for the contrasting findings as the use of holistic scores to identify prompt effect has regularly led to conclusions that prompt characteristics do not affect written product. A more detailed view of the textual features within responses can reveal significant differences that are masked when scores are used as the criterion measure for prompt effect. This study's finding that narrative responses and argumentative responses differ significantly by textual features (lexical

sophistication, syntactic complexity, and academic language use) indicates that the non-significance finding for score awarded masks the true picture of how response mode can affect written product.

8.3 Findings from the qualitative approach

In Section 8.1, the findings common to the qualitative and quantitative approaches were reported and in Section 8.2, the findings exclusive to the quantitative approach were presented. In this section the findings exclusive to the qualitative approach will be highlighted. The findings summarized below are ones that came from the stimulated recall interviews, as reported in Chapter 7.

8.3.1 Prompts with large a number of rhetorical cues

Prompts with a large number of rhetorical cues (5 or more) were viewed as easier than prompts with a small number of rhetorical cues (2 or less) by the interviewees (p.154). Five interviewees commented specifically on prompts with large numbers of rhetorical cues and stated that such prompts were easier to respond to than prompts with low numbers of rhetorical cues. This finding indicates that the number of rhetorical cues is a prompt characteristic that should be investigated further and potentially controlled for in order to help establish prompt equivalence. There should be limited variation in the number of rhetorical cues if prompts are to be equivalent and that variation should be documented in the task specifications. As this prompt characteristic was mentioned by a relatively small number of interviewees, the finding must be interpreted with caution but it is certainly a prompt characteristic worthy of future research.

The finding that test takers view writing prompts with a large number of rhetorical cues as easier than ones with low numbers is in opposition to the hypothesis put forward by Kroll & Reid (1994) who claimed that prompts with a large number of cues would be too cognitively complex for test takers. Interestingly, in this study the prompts with a large number of cues were seen as favorable by interviewees even though there was a time limit for the writing test. Because of the time constraint, interviewees felt that the large number of cues helped provide a framework for the response and that time was not wasted trying to organize the essay. This finding (though based on evidence from a limited number of test takers) potentially calls for a revisiting of the interaction between task complexity and cognitive load and the written product elicited. Kroll and Reid (1994) viewed prompts with a large number of tasks as cognitively demanding for test takers but the tentative findings from the test interviews in this thesis indicate that prompts with a large number of rhetorical cues may actually be cognitively simple for, at least test takers with advanced levels of language proficiency. Prompts with large numbers of rhetorical cues that provide a response structure to test takers without additional planning may need to be considered as reducing the cognitive complexity of the prompt, in contrast to the views of Kroll and Reid (1994).

Lim (2010) utilized number of tasks within his prompt categorization framework, with the number of tasks ranging from one to four. Lim's study did not find any significant differences in prompt difficulty based on the number of tasks, indeed the number of tasks showed the least variation of all the

categories he investigated. This finding is inconsistent with the findings from the stimulated recall interviews in this thesis, presenting an interesting contrast between evidence from test takers and the evidence provided by a quantitative analysis of task difficulty. This potentially demonstrates the need for a mixed-methods approach to investigate the effects of writing prompts, with a reliance on a single method being inadvisable.

The interviewees' comments indicate that a large number of rhetorical cues make prompts easier and this finding suggests that prompt equivalence will be easier to establish if prompts with large differences in numbers of cues are avoided.

8.3.2 Time constraint

Almost all the interviewees talked about the issues they had with completing an essay in only 30 minutes (Section 7.3.1). The interviewees reported the time constraint impacting many of the processes that were undertaken during the test; such as, selecting a prompt to respond to, planning a response, and using the prompt throughout the test. The interviewees reported using a range of test taking strategies; for example, focusing on key words in the prompt to make sure the response was on target (Section 7.3.3), planning the response before beginning to compose (Section 7.3.6), and looking back to the prompt as they composed a response to keep it relevant to the prompt (Section 7.3.5). Several of the interviewees said that they would have liked to have made more use of such strategies but felt unable to do so because of the time constraint (p.160). Of most concern is that almost all the test takers stated that 30 minutes was not sufficient for them to be able to fully demonstrate their level of writing proficiency. In all, the time constraint, along with the importance of topic familiarity were the two points made most commonly and most strongly by the interviewees.

The second language writing literature has consistently reported similar findings to those described above. Several studies (Connor & Carrell, 1993; Hale, 1992; Hall, 1991; Ruth & Murphy, 1988) have described the difficulties that second language writers have responding to a writing prompt within a limited time period. Hale's (1992) study stated the preference of test takers for 45 minutes to respond to a writing prompt, with 30 minutes being too short, according to the interviewees. The findings in the writing assessment literature are broadly comparable to those of the stimulated recall interviews. Test takers clearly feel the pressure of the time limit under test conditions and indicate that this time constraint can affect many facets of how they go about producing the final written product.

8.3.3 Importance of vocabulary

One reason interviewees expressed a preference for personal domain and open focus prompts was that they provided an opportunity to fully demonstrate the full range of the interviewees' lexical resources. The use of vocabulary was mentioned 17 times during the stimulated recall interviews (Section 7.3.7). The main focus of the interviewees' mentions of vocabulary was on the importance of selecting a prompt that was on a topic that the interviewees felt they had the lexical resources to fully respond to. Prompts that were based on topics the interviewees felt they had little relevant vocabulary for were rejected.

As was seen in Section 7.3.7, vocabulary appears to be a major factor in prompt selection and something that the interviewees consider to be important in order to perform to their best during a writing test. Interestingly, the discriminant function analysis of the discourse measures showed that the lexically focused measures (lexical sophistication and academic vocabulary use) have the two largest standardized canonical discrimination function coefficients (Table 5.9 on p.95). This shows that the two lexically focused discourse measures best indicate the significant effects of the prompt characteristics. In essence, this means that the lexical properties of the response (the frequency of the language used and its occurrence in an academic corpus) are important when it comes to determining common patterns in the responses to different prompts and distinguishing between the responses. This may suggest that test takers are right to carefully consider whether they have sufficient lexical resources to respond to a particular prompt or not, as the use of vocabulary will differentiate between responses.

To my knowledge, there is no specific mention of the importance of vocabulary in test taking strategies in the writing assessment literature. The importance of lexical knowledge to writing proficiency has been well documented (Grabe & Kaplan, 1996) but the finding here that test takers specifically consider their own lexical resources before selecting a prompt has not been commonly discussed in previous studies.

8.3.4 Importance of prompt wording

Another strategy commonly employed by the interviewees was to identify key words within the writing prompt (Section 7.3.3). These were words within the prompt that the interviewees said that they focused on while planning and/or composing their responses. The understanding derived from stimulated recall interviews was that test takers were identifying key words to help ensure that their responses were relevant to the prompt (p.153). These key words were typically identified before the test taker had begun to compose and were then referred back to regularly (within the constraints of the time limit) throughout the writing test.

Previous studies (Hoetker, 1982; Ruth & Murphy, 1984) claimed that the language proficiency of the test takers can be an important variable when constructing an understanding of the writing prompt. Misinterpretations of the prompt as a result of weak reading proficiency have been held up as a confounding variable when attempting to measure writing proficiency. However, the findings of the stimulated recall interviews do not provide any evidence that the MELAB test population has difficulty understanding the prompt wording (p.154). Powers & Fowles' (1998) interviews with GRE test takers similarly did not raise issues with specific prompt wording. While, the writing assessment literature has a history of focusing on whether test takers can successfully understand and interpret the language used in a writing prompt, to my knowledge there has been little written previously about the importance of key words in the prompt to test takers and how they use these words throughout the writing test.

It should probably be kept in mind that the MELAB test population and the GRE population is made up of relatively advanced English language users. Both tests assess college readiness and hence, are not targeting beginner or intermediate language learners. Writing tests targeting lower level language learners may need to carefully consider the difficulty of the wording used in the writing prompts. The

interviewees in this study did not have difficulty decoding any of the written language used in the prompts analyzed but this does necessarily mean that the same finding can be generalized to tests with lower proficiency populations.

8.3.5 Culturally unfamiliar topics

A lack of cultural familiarity with certain prompts was reported by five interviewees (Section 7.3.9). These comments were typically restricted to prompts that were situated in the educational domain and asked test takers to write about an aspect of the North American educational system. As the test population was largely an immigrant one, some interviewees stated that these topics were not accessible to them.

The second language writing literature has reported similar findings previously. Johns (1991) described the frustrations of an Asian student in a North American college context as he attempted to pass a high-stakes writing test. The individual student described his lack of familiarity with essay topics and how the lack of familiarity disadvantaged him from performing well on the test. He & Shi (2008) reported very similar findings for a group of Asian test takers who also reported a lack of cultural familiarity with several essay topics. The findings of Johns (1991) and He & Shi (2008) are in line with the comments made by some of the interviewees in this study. However, the findings from the stimulated recall interviews were about prompts in only a single domain (educational) and were attributable to a feature of the test population; the fact that they were generally not educated in the country where they took the test. Hence, the findings in this current work are drawn more narrowly than the broader range of topics that caused problems for the students in Johns' and He & Shi's studies.

8.4 Summary

The findings of this current work have much in common with the existing second language writing and writing assessment literatures. However, there are also some findings in this work that have been relatively little covered in the same literature. At the end of this section, a list of original findings from this study will be presented.

The common findings from the mixed-method approach are that domain and focus are two prompt characteristics that clearly contribute to problems with establishing prompt equivalence. Neither of these findings are uniquely original to this work. Qualitative research has shown previously that topic familiarity is a serious concern for takers of high-stakes writing tests. However, the specific quantitative finding that prompts in the personal domain and prompts in other domains elicit significantly different written products is a finding that is original, to the best of my knowledge. Previous studies that have quantitatively investigated prompt domain concluded that it had no significant effect on written product. Similarly, the quantitative finding that open focus prompts elicit significantly different written products is also an original contribution.

Other main findings from this thesis that have been previously reported in the literature are the effects of time constraint on test takers, the significant effect of response mode on written product, and the impact of culturally unfamiliar topics on test takers. The finding from the stimulated recall interviews that the time constraint of the MELAB writing test impacted the test taking strategies of the interviewees is echoed in the literature. However, the specific effect of narrative responses being lexically more sophisticated but syntactically more simple than argumentative responses provides a level of detail that is beyond that found in other studies. The comments from some interviewees that certain prompt topics, specifically those situated in the educational domain are culturally unfamiliar to them is quite similar to the findings of other qualitative studies of other writing test candidates. Designers of writing prompts need to be aware that some topics may be unfamiliar to certain subgroups within the test population and pretesting is an essential way to minimize the number of culturally inaccessible topics that appear on prompts on operational test forms.

One finding (although a tentative one) that is original is that a prompt with a large number of rhetorical cues is viewed by test takers as easier than a prompt with few cues. This claim has not been seen in the literature previously, to my knowledge.

The remaining findings that are original or differ from those in the existing literature are the importance of test takers' considerations of their own lexical resources when selecting a prompt to respond to, and the use of specific keywords within the prompt by test takers throughout the writing test. These lexically related findings offer something original to the writing assessment literature.

There is a recurring theme throughout the findings from both the quantitative and qualitative approaches of the importance of lexical resources within the writing assessment experience. Lexical sophistication distinguishes between different written products more than other discourse measures. Test takers are aware of the importance of their own lexical resources when they take a writing test because they talk of their desire to write on a topic that allows them to fully demonstrate their vocabulary knowledge. They know that without being able to draw on the appropriate words, they will not be able to perform well on the test. The test takers also want to focus on keywords within the prompt and use those strategically to compose a relevant and prompt-specific response. This importance of lexical competence, the ability to use low-frequency and academic vocabulary that is written in response to keywords in the prompt is a theme that is common across the mixed-methods approach of this work.

In summary, the quantitative and qualitative methods employed in this study have yielded a variety of findings, some seen previously in other studies but others that make original contributions to the second language writing and writing assessment fields. The findings that make an original contribution are:

- Prompts with large numbers of rhetorical cues are viewed as easier than prompts with only one or two cues by test takers. This is a tentative finding and will require further research to confirm.
- Prompt focus has a significant effect on the written product, with prompts that are open (that elicit a broad range of different topics) being best avoided on standardized writing assessments.

- Test takers consider their own lexical resources when selecting a prompt to respond to and will avoid prompts that they do not have sufficient vocabulary to respond to.
- Test takers focus on certain key words within the prompt and use these words repeatedly throughout the writing test. The key words are used for planning and composing and are typically used by the test taker to check that the response is relevant to the prompt.

8.5 Key findings

The discussion of the data, the prompts, and the responses above suggests that there are several prompt characteristics that may contribute to differences in the finished written products. The notable characteristics and the likely resulting textual features elicited in responses to such prompts are presented here.

1. Prompts situated in the personal domain that elicit a narrative response

As shown in Sections 8.1.1 and 8.2.1 these prompts are likely to elicit responses with high levels of lexical sophistication, low levels of academic language use and low levels of syntactic complexity.

2. Prompts situated in the personal domain that have an open focus

As shown in Sections 8.1.1 and 8.1.2 these prompts are likely to elicit responses with high levels of lexical sophistication, low levels of academic language use and low levels of syntactic complexity.

3. Prompts that elicit an argumentative response, are not situated in the personal domain, and are focused

As shown in Sections 8.1.1, 8.1.2, and 8.2.1 (and with more detail provided in Section 5.8.4) these prompts are likely to elicit responses with high levels of syntactic complexity but with low levels of lexical sophistication.

4. Prompts that elicit an argumentative response, are focused, and present academic subject matter.

As shown in Section 6.3.4 these prompts are likely to elicit responses with high levels of academic vocabulary use.

One additional prompt characteristic that appears to have a significant effect on written product is the case of prompts that contain phrasings or propositions that must be repeated in numerous sentences within the response (see Section 6.3.3). When prompts contain such phrasings or propositions (as in prompt 108), the writer is likely to need to reproduce the language in the prompt in order to write a response that is considered on topic. These prompts are likely to elicit responses with high levels of syntactic complexity.

Finally, prompts with large numbers of rhetorical cues are preferred to prompts with few rhetorical cues by test takers, although this finding is based on evidence provided by a small number of test takers. Prompts with large numbers of rhetorical cues are seen as easier than prompts with few cues because the cues provide test takers with the structure of a response and clear ideas of what content to include within the response.

Although these findings cannot be considered conclusive, due to the nesting of the prompt characteristics within the research design, the findings are indicative that there are specific writing prompt characteristics that may result in significant differences in written product. The MANOVA data showed that there were no significant differences in writing scores awarded to test takers regardless of the prompt they selected, although the use of holistic scores on the MELAB Writing Test may mask the differences in written product revealed by the MANOVA findings. The use of analytic scores may have revealed observable differences in the quality of the written responses. The fact that significant differences in important textual features were not reflected in the scores raises some serious questions about the use of holistic scores on high-stakes exams and whether an alternative approach to scoring (for example, analytic scoring) may be superior. If holistic scores are the only ones reported, the extent to which test users are provided with meaningful and actionable score information on test takers' writing skills is open to question.

It is clear that there are a range of significant differences in key textual features of the responses depending on which prompt was selected. This situation creates some issues with score interpretation. If the same score may be awarded to responses with quite different textual profiles, how can users of the test scores interpret what the scores mean in terms of the writing proficiency of the test taker? It is this issue of score interpretation that seems to be the main concern raised by the findings reported above. This concern will be discussed further in Chapter 9.

Chapter 9 – Conclusion

This study set out to investigate the relationships and interactions between test takers, independent writing prompts, and the final written product during a high-stakes ESL test. The main aims of the work were to identify a set of core independent prompt characteristics that could consistently distinguish between different prompts and to then investigate how test taking processes and written product were affected if those prompt characteristics were manipulated. The review of literature yielded three research questions.

1. What are the distinguishing characteristics of independent writing prompts?
2. How do these characteristics affect the test-takers' final written product?
3. How do these characteristics affect the test takers' test taking processes?

To answer the first research question, a triangulated approach to analyzing independent writing prompts was taken. A taxonomy of previous studies of prompt categorization from the literature was applied to a large number of independent writing prompts. In addition, experienced essay raters were asked to read responses to several prompts that had been highlighted as potentially eliciting different written products. Finally, test takers were interviewed to explore the prompt characteristics that were seen as important during the test taking process. The synthesis of these findings was reported in Chapter 4 and produced an answer to the first research question. The distinguishing characteristics of independent writing prompts are:

- Domain (personal, public, occupational, educational)
- Response mode (narrative or argumentative)
- Number of rhetorical cues
- Focus (open or focused)

To answer the second research question, 60 essays were collected in response to each of six independent writing prompts. These prompts differed based on the above prompt characteristics. The sample of 360 essays was drawn from live test administrations and written by a sample population diverse in gender, language background, and cultural background. The 360 essays were analyzed using numerous discourse measures to operationalize important traits of second language writing proficiency such as, syntactic complexity, lexical sophistication, cohesion, accuracy, and fluency. Statistical analyses (principal component analysis, discriminant function analysis, and MANOVA) applied to the dataset of discourse measure data, and described in Chapter 5 produced an answer to the second research question. There were significant effects of the independent writing prompt characteristics on:

- Lexical sophistication
- Academic vocabulary use
- Syntactic complexity

There were also significant effects of the prompt characteristics on cohesion and fluency but the effect sizes were smaller than for the three textual features listed above. There was no significant effect of the independent prompt characteristics on the accuracy of the written language produced or on the holistic scores awarded. Table 9.1 below presents a summary of the main relationships between prompt characteristics and textual features of the responses.

Table 9.1: Relationships between prompt characteristics and textual features of the responses

Prompt characteristic(s)	**Textual features of response**
Personal domain, narrative response	High levels of lexical sophistication Low levels of academic vocabulary use Low levels of syntactic complexity
Personal domain, open focus	High levels of lexical sophistication Low levels of academic vocabulary use Low levels of syntactic complexity
Argumentative response, non-personal domain, focused	High levels of syntactic complexity Low levels of lexical sophistication
Argumentative response, focused, academic subject matter in public or educational domains	High levels of academic vocabulary use

Also, prompts that contained phrases or propositions that were reproduced repeatedly in the response elicited responses with high levels of syntactic complexity. These findings answer the second research question and demonstrate how the prompt characteristics affect the written product.

To answer the third and final research question, 28 test takers underwent two phases of stimulated recall interviews. The interviews focused on how test takers selected a prompt to respond to, planned a response, and then composed a response. The interview recordings were transcribed and then analyzed for common themes regarding test taking processes. The analyses of the interview transcripts, reported in Chapter 7 produced an answer to the third research question. The prompt characteristics of domain, number of rhetorical cues, and focus were shown to have an effect on test takers' choice of prompt. Test takers reported that the following prompt characteristics contributed to prompts that were relatively easy to respond to:

- Prompts that were based on familiar topics, typically those situated in personal or occupational domains.
- Prompts that had an open focus and allowed test takers to respond on a broad range of different topics.
- Prompts that had a large number of rhetorical cues.

The interviews revealed that the prompt was then used to plan and structure a response, and the prompt was referred back to by the test taker while composing. In all, three of the core prompt characteristics (domain, focus, and number of rhetorical cues) had a range of effects on test taking processes, including prompt selection, response planning, and the composing of the response.

The study as a whole shows that it is possible to identify a common set of prompt characteristics that help to distinguish between different independent writing prompts. A mixed-methods approach to

investigate the effect of these characteristics on both written products and test taking processes reveals that both the test taker interactions with the prompt and the written language produced in response to the prompt vary based on the characteristics of the prompt. However, the accuracy of the responses and the holistic writing scores awarded to the responses do not vary significantly, regardless of the prompt characteristics. Hence, the main concern resulting from these findings is one of score interpretation. While the score awarded to a test taker may not vary based on the prompt, the nature of the written language elicited may vary significantly. This may be of particular concern to the users of test scores, such as institutions of higher learning, employers, and authorities that make immigration decisions.

Responses to some prompts may tend to be lexically rich but syntactically simple. However, responses to other prompts may tend to be syntactically complex and lexically simple. If these very different types of written responses have been awarded the same score, how is that score to be interpreted? What inferences may be drawn about the test taker's second language writing proficiency if her or his score may reflect different abilities from that of another test taker, depending on the prompt and not on real differences in writing ability? It is these issues of score interpretation that arise from the findings of this study. The question of whether different prompts can be considered equivalent needs to be considered in light of whether issues of score interpretation are a serious concern or not. If the only concern is the score awarded, then writing prompts with quite different characteristics may be considered equivalent as the score awarded (at least if it is a holistic score) is unlikely to vary significantly regardless of the prompt.

However, if the score is to be interpreted in terms of the competencies that the score represents, then this will present potential issues with establishing prompt equivalence. As different prompts with different characteristics elicit written language with varying textual profiles, the prompts cannot be considered equivalent if detailed inferences about writing proficiency need to be drawn from the score. For example, if any diagnostic information about a test taker's writing proficiency is required, such as the writer's strengths and weaknesses with grammatical control, lexical competence, or ability to write in the academic domain rather than the personal, then this study has suggested that writing prompts that vary by the prompt characteristics explored here cannot be considered equivalent. The prompt characteristics found to have a significant effect on the textual features of the response must be controlled for, if the prompts are to be considered equivalent and meaningful for a specific context, and if stable inferences are to be drawn from the scores.

An additional concern is the relationship between construct coverage, prompt characteristics, and the resulting textual features of the responses. These important aspects of the assessment need to be aligned. If only a single writing prompt is administered and the prompt characteristics are controlled, in line with the findings of this thesis, then both the prompts and the textual features of the responses will be quite narrowly drawn. In this situation, there is the risk that the writing construct may be under represented. An alignment between the test construct, prompt characteristics, and textual features of the responses may well be challenging to achieve with only a single writing prompt administered on the assessment. While a refined understanding of the relationships between prompt characteristics and textual features of the responses is valuable, this understanding should be used to aid test design and to help ensure construct coverage., in terms of the prompts presented on the assessment and the written language elicited by these prompts.

It should be emphasized that the findings from this study should not be considered as, nor are they claimed to be definitive. The findings are only indicative as the nesting of the prompt characteristics prevents definitive relationships between individual prompt characteristics and specific textual features of the responses from being isolated. This weakness in the research design does not mean the findings

cannot be generalized, however the findings do provide only an indication of the specific relationships between prompt characteristics and the textual features of the responses. As has been noted previously (p.64), the nesting within the research design was unavoidable due to the preference for collecting responses to prompts administered during live test administrations. This choice ensured that the responses had been produced by writers who were motivated to fully demonstrate their true level of writing proficiency. Collecting all data during live test administrations was seen as more important than employing an experimental design which would have allowed for the issue of nesting to be overcome. An additional limitation is that incidences of high academic vocabulary use may be partly attributable to academic language having been lifted from the prompt wording. A further investigation of solely original language elicited from prompts would be needed to confirm the effect of prompt characteristics that elicit responses with high levels of academic language use.

The findings of this study indicate that there are a set of writing prompt characteristics that can have a significant effect on the written products and on test taking processes. These prompt characteristics do not have a significant effect on the holistic score awarded but they do have a significant effect on the textual features of the response. The main concerns arising from this conclusion are with regard to score interpretation and construct coverage. The inferences drawn from a score about second language writing proficiency may differ substantively depending on the characteristics of the prompt. This is the context in which prompt equivalence must be considered, along with an appreciation of the writing construct to be covered and the writing prompt(s) needed to represent that construct.

If the score itself is all that is of interest to stakeholders, then tasks may be considered equivalent even if they differ by the prompt characteristics described in this thesis. However, if the inferences drawn from the score are considered important and if any diagnostic information regarding test takers' writing proficiency needs to be gleaned from the score, then writing prompts that differ by distinguishing prompt characteristics cannot be considered equivalent. It is the question of how scores are to be used that will determine whether prompts may be considered equivalent or not. This refined understanding of prompt equivalence will then help inform the number and types of prompts required to adequately cover the construct definition of the assessment.

9.1 Implications

There are implications for the designers and developers of second language writing tests from the findings reported above. The primary implication cover three main areas:

- Construct definition
- Task specifications
- Rating scale design

9.1.1 Construct definition

The findings of this thesis indicate that writing prompts situated in different domains and that elicit different response modes can elicit responses that differ significantly by textual profile. The focus of the prompt and the number of rhetorical cues can also contribute to differences in textual profile. These findings raise the question of how the writing construct is to be defined for a given assessment. If the construct of writing proficiency to be measured is drawn very broadly, as simply a measure of general writing proficiency, or perhaps, as simply a measure of academic writing proficiency it will be challenging to precisely define the attributes of the test input (the prompt) that allows construct relevant written language to be elicited. The nature of the written language that aligns with the construct definition must

be considered carefully and only after these decisions have been made can the input attributes be defined, as the findings of this work indicate that there is a close relationship between prompt characteristics (an attribute of the test input) and the textual profile of the response. An additional consideration will be the number of prompts that are required to elicit the written language (of varying textual profiles) required to ensure construct coverage. With a refined understanding of the likely textual profiles of language elicited by particular prompt types, the number of prompts required and the different prompt characteristics that must be represented (to elicit the required language) should be easier to predict.

If the construct can be drawn more narrowly; for example, if a writing test is designed to measure the ability to write a personal narrative, then the construct definition may state specifically the characteristics of the prompt that will allow the type of written language desired to be elicited. Of course, if the construct is not defined narrowly, then the number of writing prompts needed to guarantee construct coverage will need to be carefully determined. But without a clear construct definition, other test related documentation, such as task specifications and rating scales will be difficult to design with sufficient specificity to ensure that the construct is fully covered and represented throughout the assessment process. A consideration of the domain(s) and response mode(s) that are to be targeted, along with a consideration of other attributes of the input, such as the focus of the prompt and the number of rhetorical cues will allow for a more focused and specific construct definition. Ultimately, this will also allow for the resulting scores awarded to the written product to be interpreted more straightforwardly and precisely, as those making decisions based on the score and the inferences drawn from the score will have a clearer idea of what the score means and which inferences can be appropriately drawn from that score.

This issue of construct definition may be of particular concern to gatekeeping tests such at TOEFL, IELTS, and MELAB, which are used to assess readiness for admission to English language medium universities. If these tests do not measure the ability to write in an academic context, the resulting scores are of limited use to admissions officers who use test scores to determine readiness to function in English in an academic environment. The construct of writing measured on English language international admissions tests needs to be aligned with the type of academic texts required at university. Without a clear construct definition, based on the ability to produce writing in an academic context, the scores of the writing component of these admissions tests will be of limited worth.

As stated previously (p.137), it is challenging to see the relevance of the evidence provided by narrative responses in the personal domain to the construct of the ability to write in academic (or even professional) contexts. The MELAB Writing construct definition is quite broad (see Section 4.2.1), requiring both formal and casual communication contexts within the educational and professional domains to be assessed. Given this broad construct definition, it is challenging to see how a single writing prompt can elicit a sufficiently broad range of written language to sufficiently cover the construct. This leads to the conclusion that the MELAB Writing section would be better served with at least two writing prompts for students to respond to, instead of a single prompt.

9.1.2 Task specifications

The findings of this study indicate that there are four distinguishing prompt characteristics that contribute to prompt equivalence. These are domain, response mode, focus, and number of rhetorical cues. The findings also indicate that there are particular aspects of these prompt characteristics that are most likely to introduce problems with prompt equivalence. These are:

- Prompts situated in the personal domain
- Prompts that elicit a narrative response
- Prompts with an open focus
- Prompts with a large number of rhetorical cues

If the test construct is defined so that these prompt characteristics are not desired within a particular assessment context, then task specifications can be crafted so that these characteristics do not feature in any prompt within an operational test form.

Typically, the construct definition presents a theoretical argument for and justification of what is to be measured on a particular component of an examination. The construct definition is commonly tied to the findings in the relevant academic literature. The construct is a theoretical justification for and description of what the test intends to measure. The task specifications, however present a less academic and more mechanical set of instructions to test developers and reviewers. The task specifications set out a blueprint or recipe for developers and reviewers to follow. The task specifications present clear and detailed guidance as to how the item, testlet, or prompt should be designed, often with a great deal of detail; such as word count, readability indices, level of targeted syntactic complexity, word frequency, and so on, depending on the item type in question. Task specifications may also detail any of the prompt characteristics reported in this study. The more detailed the guidance to test developers and reviewers provided in these specifications, the better the likelihood that prompts can be consistently developed and administered across different test forms that are equivalent in terms of the challenge and opportunity they present test takers to demonstrate their true level of writing proficiency. Hence, the ability to create more detailed task specifications, featuring guidance on the prompt characteristics of domain, response mode, focus, and the number of rhetorical cues will be possible and these specifications should be clearly related to the construct definition.

9.1.3 Rating scale

Finally, one additional piece of documentation that may be refined on the basis of the findings of this thesis is the rating scale. Rating scales apply certain scoring criteria, such as accuracy, vocabulary use, development of ideas, and so on to evaluate the written product according to a number of different levels of defined performance. The scale can then be used by raters to assign either holistic or analytic scores to a written response. There are several different approaches to designing rating scales described in the literature but most approaches require the justification of the scoring criteria that serve as the basis for how the responses will be evaluated. This justification of the scoring criteria should come, at least in part from the academic literature and also from the construct definition. If the construct has been defined narrowly, as described above and takes into account characteristics such as domain and response mode, then the textual features that can be expected in the responses are more predictable. If the construct definition includes the ability of test takers to write a personal narrative, the findings of this work indicate that lexically sophisticated but syntactically simple responses can be expected. In this case, the rating scale should be designed so the scoring criteria and band level descriptors are crafted so as to capture the expected textual features of the responses. That is, the rating scale can be more relevant to the textual features typically elicited if the construct has been narrowly defined and the prompt characteristics that are permissible have been explicated in the task specifications.

The MELAB Writing rating scale may be improved with the addition of descriptors that address specific textual features shown to be important in this thesis. Descriptors that address the use of academic vocabulary would allow raters to better distinguish between the use of low frequency vocabulary and

lexical choices that are relevant to the academic writing context. This would be an especially positive change for an assessment that is often taken for academic purposes. Additionally, a specific mention of fluency, or the length of the response may also be considered. Finally, the reporting of holistic scores is open to question. Analytic scores that provide more nuanced information about test takers ability to write syntactically complex, lexically rich, cohesive, fluent responses would be a positive change and provide both test takers and other stakeholders with a more complete picture of the second language writing skills represented by the writing score.

The findings of this thesis can contribute to better designed second language writing assessments by improving the understanding of the relationships between prompt characteristics and textual features exhibited in responses to those prompts. The findings can contribute to more precisely defining the construct of writing to be assessed, tightening the level of detail described in the task specifications, and refining the scoring criteria and descriptor wording in the rating scale. These changes would potentially allow stakeholders to draw clearer and more precise inferences from the scores awarded to test takers.

9.2 Future research directions

This thesis has shown that the second language writing assessment field would benefit from further qualitative and quantitative research that focuses on how test takers engage with the writing test environment. For example, the field would greatly benefit from a better understanding of the appropriate length of time that should be provided for test takers, especially for high-stakes tests. Investigating how test takers respond to different lengths of time to complete a writing assessment would be very helpful. It would be particularly interesting to explore whether there are significant differences in written product (both textual features and scores awarded) as a result of test takers being allowed different lengths of time to complete the test.

The findings of this study also indicate that further qualitative research that engages directly with the test taker via interviews or focus groups would provide a clearer insight into how test takers engage with and respond to writing prompts. The voice of the test taker is seldom heard in the writing assessment literature. Such research, with, for example, a focus on how test takers engage with integrated (reading-to-write or listening-to-write) prompts would be very valuable for broadening our understanding of the effect of integrated prompts on test takers and test performance.

With regard to the prompt characteristics that were studied in this research, more detailed explorations of the effect of domain and response mode on written product would be of interest. An investigation of prompts that differed only by domain would allow for a more conclusive understanding of the effect of that prompt characteristic. Prompts in the personal domain had a significant effect on the written product in this study but the stimulated recall interview findings indicated that any prompt with familiar topics would be favored by test takers. Prompts based on familiar topics in other domains (not personal) may also have an effect on prompt equivalence. Research that can offer conclusive evidence on the relationship between domain and the effect on written product and test taking processes would be helpful. Similarly, a study that focuses on response mode and incorporates prompts that elicit expository responses, in addition to narrative and argumentative responses would be of interest. This thesis showed that narrative responses and argumentative responses differed significantly by textual features. However, the nesting within the prompt design meant that it was difficult to be conclusive regarding the contribution of the response mode to these differences. A study that isolated only response mode, and brought in expository responses as part of the design would allow for more conclusive findings on this important prompt characteristic.

In conclusion, the findings of this study indicate that establishing prompt equivalence within a testing program will be challenging if many different prompts are used across administrations. The main concern is not the consistency of the scores that are awarded but how the scores may be interpreted. If the issue of score interpretation is of concern, then prompts can only be considered equivalent if the prompt characteristics investigated in this study are controlled for. This is a serious issues for test design and validation, and one that warrants further research that will ultimately benefit the test takers and other stakeholders.

Appendix 1

Prompts used in Greenberg (1981).

High cognitive / low experiential demands

1. In most American colleges, students must pass required courses in English, math, and science before they are allowed to take courses in their major area of study. Instead of making all students attend all of their required courses, colleges should offer more independent study programs in which students could complete some of their courses on their own, working at their own pace.

Do you agree or disagree with this statement? In an essay of about 300 words, explain and illustrate your answer in detail.

2. The traditional grading system used by most colleges includes A, B, C, D, and F grades. This should be changed to a Pass/Fail system for all of your courses because the traditional grading system forces you to compete for good grades rather than to strive for your own knowledge and learning.

Do you agree or disagree with this statement? In an essay of about 300 words, explain and illustrate your answer in detail.

High cognitive / high experiential demands

1. In most American colleges, students must pass required courses in English, math, and science before they are allowed to take courses in their major area of study. Instead of making all of you attend all your required courses, colleges should offer you more independent study programs in which you could complete some of these courses on your own, working at your own pace.

Do you agree or disagree with this statement? In an essay of about 300 words, explain and illustrate your answer in detail.

2. The traditional grading system used by most colleges includes A, B, C, D, and F grades. This should be changed to a Pass/Fail system for all of your courses because the traditional grading system forces you to compete for good grades rather than to strive for your own knowledge and learning.

Do you agree or disagree with this statement? In an essay of about 300 words, explain and illustrate your answer in detail.

Low cognitive / low experiential demands

1. In most American colleges, students must pass required courses in English, math, and science before they are allowed to take courses in their major area of study. Some colleges, however, offer students the option of taking independent study programs in which they can complete some of these required courses on their own, working at their own pace.

In an essay of about 300 words, discuss independent study as an alternative to attendance at required courses. In your essay, you might wish to choose one of the following strategies: contrast taking an independent study program to attending a required course, or explain some reasons for taking required courses, or suggest some change in the rules concerning required courses and give some reasons for these changes.

2. The traditional grading system used by most colleges includes A, B, C, D, and F grades. Some colleges however, use a Pass/Fail system for all courses in order to move the focus of an education away from getting a certain grade to striving for knowledge and learning.

In an essay of about 300 words, discuss the traditional grading system (A-F). In your essay, you might wish to choose one of the following strategies: contrast the traditional grading system to a different type of grading system, or explain some reasons for keeping the traditional grading system, or suggest some changes in the traditional grading system and give reasons for these changes.

<u>Low cognitive / high experiential demands</u>

1. In most American colleges, students must pass required courses in English, math, and science before they are allowed to take courses in their major area of study. Some colleges, however, offer you the option of taking independent study programs in which you can complete some of these required courses on your own, working at your own pace.

In an essay of about 300 words, discuss independent study as an alternative to attendance at your required courses. In your essay, you might wish to choose one of the following strategies: contrast taking an independent study program to attending a required course, or explain some reasons for taking required courses, or suggest some changes in the rules concerning required courses and give reasons for these changes based on your experience.

2. The traditional grading system used by most colleges includes A, B, C, D, and F grades. Some colleges, however, use a Pass/Fail system for all courses in order to move the focus of your education away from getting a certain grade to striving for knowledge and learning in your classes.

In an essay of about 300 words, discuss the traditional grading system (A-F). In your essay, you might wish to choose one of the following strategies: contrast the traditional grading system to a different type that you have experienced, or explain some reasons for keeping the traditional grading system, or suggest some changes in the traditional grading system and give reasons for these changes based on your experiences.

Appendix 2

MELAB Composition Descriptions

97 Topic is richly and fully developed. Flexible use of a wide range of syntactic (sentence-level) structures, and accurate morphological (word forms) control. Organization is appropriate and effective, and there is excellent control of connection. There is a wide range of appropriately used vocabulary. Spelling and punctuation appear error-free.

93 Topic is fully and complexly developed. Flexible use of a wide range of syntactic structures. Morphological control is nearly always accurate. Organization is well controlled and appropriate to the material, and the writing is well connected. Vocabulary is broad and appropriately used. Spelling and punctuation errors are not distracting.

87 Topic is well developed, with acknowledgment of its complexity. Varied syntactic structures are used with some flexibility, and there is good morphological control. Organization is controlled and generally appropriate to the material, and there are few problems with connection. Vocabulary is broad and usually used appropriately. Spelling and punctuation errors are not distracting.

83 Topic is generally clearly and completely developed, with at least some acknowledgement of its complexity. Both simple and complex syntactic structures are generally adequately used; there is adequate morphological control. Organization is controlled and shows some appropriacy to the material, and connection is usually adequate. Vocabulary use shows some flexibility, and is usually appropriate. Spelling and punctuation errors are sometimes distracting.

77 Topic is developed clearly but not completely and without acknowledging its complexity. Both simple and complex syntactic structures are present; in some 77 essays, these are cautiously and accurately used while in others there is more fluency and less accuracy. Morphological control is inconsistent. Organization is generally controlled, while connection is sometimes absent or unsuccessful. Vocabulary is adequate but may sometimes be inappropriately used. Spelling and punctuation errors are sometimes distracting.

73 Topic development is present, although limited by incompleteness, lack of clarity, or lack of focus. The topic may be treated as though it has only one dimension, or only one point of view is possible. In some 73 essays, both simple and complex syntactic structures are present, but with many errors; others have accurate syntax but are very restricted in the range of language attempted. Morphological control is inconsistent. Organization is partially controlled, while connection is often absent or unsuccessful. Vocabulary is sometimes inadequate, and sometimes inappropriately used. Spelling and punctuation errors are sometimes distracting.

67 Topic development is present but restricted, and often incomplete or unclear. Simple syntactic structures dominate, with many errors; complex syntactic structures, if present, are not controlled. Lacks morphological control. Organization, when apparent, is poorly controlled, and little or no connection is apparent. Narrow and simple vocabulary usually approximates meaning but is often inappropriately used. Spelling and punctuation errors are often distracting.

63 Contains little sign of topic development. Simple syntactic structures are present, but with many errors; lacks morphological control. There is little or no organization, and no connection apparent.

Narrow and simple vocabulary inhibits communication, and spelling and punctuation errors often cause serious interference.

57 Often extremely short; contains only fragmentary communication about the topic. There is little syntactic or morphological control, and no organization or connection are apparent. Vocabulary is highly restricted and inaccurately used. Spelling is often indecipherable and punctuation is missing or appears random.

53 Extremely short, usually about 40 words or less; communicates nothing, and is often copied directly from the prompt. There is little sign of syntactic or morphological control, and no apparent organization or connection. Vocabulary is extremely restricted and repetitively used. Spelling is often indecipherable and punctuation is missing or appears random.

0 *A zero can be given to a nonresponse.* A completely blank answer sheet or simply the test taker's name on the space where the essay should be written.

A zero can also be given to a composition that is written on a topic different from any of those assigned. Connection of composition to the prompt may be so loose that the essay could very well have been prepared in advance. Considerable effort must be made to see the connection between the composition and the prompt.

Appendix 3

Performing error counts

In order to ensure that error counts were performed consistently, the error count guidelines recommended by Polio (1997:139) were adopted. In an effort to count the number of errors objectively, four experienced MELAB essay raters (including the researcher) read Polio's article and error guidelines and then each read the same five essays independently. The raters then met to clarify some differences, particularly in how errors were counted for collocation and stylistic choices. One of the raters tended to identify irregular stylistic choices or collocations that were slightly non-native like as errors while the other raters were not consistently counting such cases as errors. Through discussion the raters, as a group determined that the one rater was being excessively harsh. Following this meeting, each of the four raters independently counted errors in a further 25 essays.

The correlation coefficients for the four raters ranged from a low of .861 to a high of .984. The coefficients indicate a high rate of agreement between the raters in terms of their consistency in counting errors in line with Polio's guidelines. The correlations achieved between the raters are in line with those reported by Polio (1997: 128) of between .89 and .94. The rater correlations were considered to be sufficiently high to indicate that the author was analyzing essays for error in a systematic way and consistent with the principles set out by Polio. The author then counted and recorded the errors in the remaining 330 essays personally.

Both the total number of errors in a responses were recorded along with an error count that is standardized for response length (number of errors per 100 words). Recording both counts (total and standardized) helps to replicate how MELAB raters are trained and apply the MELAB rating scale when scoring live MELAB responses. Raters read for both error quantity and error severity, interpreting how distracting or confusing the errors are to the readability of the response. The total number of errors within a response, along with their severity will influence the score awarded to the response. However, the length of the response will interact with the quantity and severity of the error and influence the evaluation. Long responses that have a sustained error rate may make a more negative impression than a shorter response with a similar rate of error. Conversely, a very simplistic and short response that has a high error rate despite the relative lack of ambition shown by the writer, may be evaluated more harshly than a response that is lengthy and attempts many complex constructions. This interplay between the length of the response and the error rate is the reason for recording the two accuracy measures.

Appendix 4

Correlation matrix between the latent variables

Correlation Matrix[a]

		word #	ASL	AWL	SYNLE	CONi	CONCAUSi	CONLOGi	CONTEMPi	TTR	FREQ 1	FREQ 2	FREQ 3	FREQ AC	TOTERR	ERR/100
Correlation	word #	1.000	.093	-.080	.090	.008	-.049	-.001	-.054	-.101	.067	-.016	-.099	-.095	.310	-.210
	ASL	.093	1.000	-.020	.508	.271	.003	.178	-.017	.009	.165	-.081	-.187	.048	-.015	-.082
	AWL	-.080	-.020	1.000	-.026	-.064	-.087	-.017	-.107	.203	-.324	.290	.224	.684	-.045	-.002
	SYNLE	.090	.508	-.026	1.000	.210	-.003	.168	.061	-.010	.072	-.035	-.080	.080	.062	-.042
	CONi	.008	.271	-.064	.210	1.000	.458	.634	.355	-.077	.074	-.077	-.025	-.043	.060	.037
	CONCAUSi	-.049	.003	-.087	-.003	.458	1.000	.571	-.146	-.262	.269	-.245	-.192	-.101	.105	.111
	CONLOGi	-.001	.178	-.017	.168	.634	.571	1.000	-.012	-.229	.285	-.254	-.200	-.006	.112	.092
	CONTEMPi	-.054	-.017	-.107	.061	.355	-.146	-.012	1.000	.147	-.144	.118	.137	-.053	-.086	-.036
	TTR	-.101	.009	.203	-.010	-.077	-.262	-.229	.147	1.000	-.389	.353	.276	.182	-.229	-.175
	FREQ 1	.067	.165	-.324	.072	.074	.269	.285	-.144	-.389	1.000	-.809	-.780	-.307	.114	.056
	FREQ 2	-.016	-.081	.290	-.035	-.077	-.245	-.254	.118	.353	-.809	1.000	.279	.257	-.099	-.060
	FREQ 3	-.099	-.187	.224	-.080	-.025	-.192	-.200	.137	.276	-.780	.279	1.000	.235	-.085	-.029
	FREQ AC	-.095	.048	.684	.080	-.043	-.101	-.006	-.053	.182	-.307	.257	.235	1.000	-.110	-.056
	TOTERR	.310	-.015	-.045	.062	.060	.105	.112	-.086	-.229	.114	-.099	-.085	-.110	1.000	.820
	ERR/100	-.210	-.082	-.002	-.042	.037	.111	.092	-.036	-.175	.056	-.060	-.029	-.056	.820	1.000
Sig. (1-tailed)	word #		.038	.064	.044	.441	.178	.490	.154	.028	.102	.382	.030	.036	.000	.000
	ASL	.038		.356	.000	.000	.476	.000	.377	.435	.001	.062	.000	.181	.389	.059
	AWL	.064	.356		.314	.114	.049	.371	.021	.000	.000	.000	.000	.000	.198	.488
	SYNLE	.044	.000	.314		.000	.475	.001	.125	.427	.087	.254	.064	.066	.119	.213
	CONi	.441	.000	.114	.000		.000	.000	.000	.073	.082	.071	.320	.207	.127	.242
	CONCAUSi	.178	.476	.049	.475	.000		.000	.003	.000	.000	.000	.000	.028	.024	.018
	CONLOGi	.490	.000	.371	.001	.000	.000		.412	.000	.000	.000	.000	.452	.017	.041
	CONTEMPi	.154	.377	.021	.125	.000	.003	.412		.003	.003	.013	.005	.156	.051	.245
	TTR	.028	.435	.000	.427	.073	.000	.000	.003		.000	.000	.000	.000	.000	.000
	FREQ 1	.102	.001	.000	.087	.082	.000	.000	.003	.000		.000	.000	.000	.015	.143
	FREQ 2	.382	.062	.000	.254	.071	.000	.000	.013	.000	.000		.000	.000	.030	.127
	FREQ 3	.030	.000	.000	.064	.320	.000	.000	.005	.000	.000	.000		.000	.054	.291
	FREQ AC	.036	.181	.000	.066	.207	.028	.452	.156	.000	.000	.000	.000		.019	.146
	TOTERR	.000	.389	.198	.119	.127	.024	.017	.051	.000	.015	.030	.054	.019		.000
	ERR/100	.000	.059	.488	.213	.242	.018	.041	.245	.000	.143	.127	.291	.146	.000	

Appendix 5

PARTICIPANT CONSENT FORM

Project title ……

Researcher's name ……

Supervisor's name ……

- I have read the Participant Information Sheet and the nature and purpose of the research project has been explained to me. I understand and agree to take part.
- I understand the purpose of the research project and my involvement in it.
- I understand that I may withdraw from the research project at any stage and that this will not affect my status now or in the future.
- I understand that while information gained during the study may be published, I will not be identified and my personal results will remain confidential. {*If other arrangements have been agreed in relation to identification of research participants this point will require amendment to accurately reflect those arrangements*}
- I understand that I will be audiotaped / videotaped during the interview. {*Omit this point if the interview will not be taped}*
- I understand that data will be stored … {*insert details of how and where data – including hard and electronic copies of transcripts, or any video or audiotapes used – will be stored, who will have access to it and what limits will be placed on that access}*
- I understand that I may contact the researcher or supervisor if I require further information about the research, and that I may contact the Research Ethics Coordinator of the School of Eduction, University of Nottingham, if I wish to make a complaint relating to my involvement in the research.

Signed …… (research participant)

Print name …… **Date** ……

Contact details

Researcher: *{complete preferred contact details}*

Supervisor: *{complete preferred contact details}*

School of Education Research Ethics Coordinator: educationresearchethics@nottingham.ac.uk

2011/18/RM

School of Education – Research Ethics Approval Form

The University of Nottingham

Name	Mark Chapman
Main Supervisor	Liz Hamp-Lyons
Course of Study	PhD
Title of Research Project:	Prompt equivalence in second language writing assessment
Is this a resubmission?	No

Date statement of research ethics received by PGR Office: 16/05/11

Research Ethics Coordinator Comments:

Dear Mark

This all seems fine to me and I am happy to grant you ethical approval to proceed on the basis set out in this submission. I personally would be inclined to simply the project description that you give to participants. I am happy to leave you to discuss that with your supervisors.

Good luck,

Roger

I consider this research to be above minimum risk ☐

Outcome:
Approved ✓ Revise and Resubmit ☐

Signed: Name: Prof Roger Murphy (Research Ethics Coordinator) Date: 27.05.11

Appendix 6

For office use only			
1		R1	
2		R2	
3		R3	

Name: (print) _____ Date: _____
Family/Last Given/First Name

Signature: _____ ID Number: _____

MELAB
Michigan English Language Assessment Battery

Writing Section

Topic Set **2012 Sample Test**

Instructions

1. You will have **30 minutes** to write on one of the two topics printed on the next page. If you do not write on one of these topics, your test will not be scored.
2. You may make an outline if you wish, but your outline will not count toward your score.
3. Write about 1 to 2 pages. Your test will be marked down if it is extremely short. Ask the examiner for more paper if you need it.
4. You will not be graded on the appearance of your test, but your handwriting must be readable. You may change or correct your writing, but you should not copy the whole composition over.
5. Your test will be judged on clarity and overall effectiveness, as well as on:
   - topic development
   - organization
   - range, accuracy, and appropriateness of grammar and vocabulary
6. You should write on **only one** of the two topics. **Circle** the letter (A or B) of the topic you choose.

Do not turn the page until the examiner tells you to.

Write on **only one** of the two topics.

Remember to **circle** the letter (A or B) of the topic you choose.

Topic Set: 2012 Sample Test

A. The use of computers in elementary school classrooms is becoming very widespread. What positive and negative effects do you think this has on student learning? Support your views with specific examples.

B. When visiting a new city or country, some people like to go on group tours led by professional tour guides. Others prefer to explore new places on their own. Which do you prefer? Give reasons and examples to support your opinion.

References

Adkins, D. (1947). *Construction and Analysis of Achievement Tests*. Washington D.C.: US Government Printing Office.

Anderson, C. C. (1960). The New Step Essay Test as a Measure of Composition Ability. *Educational and Psychological Measurement* 20, 95-102.

Ballard, P. B. (1925). *The New Examiner*. London: Hodder and Stoughton.

Banerjee, J., Yan, X., Chapman, M., & Elliott, H. (2015). Keeping up with the times: Revising and refreshing a rating scale. *Assessing Writing*, 26, 5-19.

Barkaoui, K. (2007). Participants, texts and processes in ESL/EFL essay tests: A narrative review of the literature. *The Canadian Modern Language Review*, 64(1), 99-134.

Barkaoui, K., Brooks, L., Swain, M., & Lapkin, S. (2013)**.** Test-takers' strategic behaviors in independent and integrated speaking tasks. *Applied Linguistics*, 34(3), 304-324.

Barton, E. (2004) Linguistic discourse analysis: how the language in texts works. In C. Bazerman & P. Prior (Eds.). *What writing does and how it does it* (pp. 57-82). Mahwah, NJ: Lawrence Erlbaum Associates.

Biber, D. (1992). On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes*, 15 (2), 133-163.

Braddock, R., Lloyd-Jones, R., & Schoer, L. (1963). *Research in Written Composition*. Urbana, IL: National Council of Teachers of English.

Breland, H., Lee, Y., Najarian, M. & Muraki, E. (2004). *An analysis of TOEFL CBT writing prompt difficulty and comparability for different gender groups* (ETS TOEFL Research Report 76). Princeton, NJ: Educational Testing Service

Bridgeman, B. & Carlson, S. (1983)*. Survey of Academic Writing Tasks Required of Graduate and Undergraduate Foreign Students*. ETS RR No. 83-18. Princeton, NJ: Educational Testing Service.

Brossell, G. (1983). Rhetorical specification in essay examination topics. *College English*, 45(2), 165-173.

Brossell, G. & Ash, B. (1984). An experiment with the wording of essay topics. *College Composition and Communication*, 35(4), 423-425.

Brown, A. (2007). An investigation of the rating process in the IELTS oral interview. *In Taylor, L. & Falvey, P. (Eds.) IELTS Collected Papers: Research in Speaking and Writing Assessment (pp. 379-421). University of Cambridge Local Examinations Syndicate.*

Brown, J. D. (1988). *Understanding research in second language learning: A teacher's guide to statistics and research design*. Cambridge University Press.

Brown, J.D., Hilgers, T., & Marsella, J. (1991). Essay prompts and topics: Minimizing the effects of mean differences. *Written Communication* 8(4), 533-556.

Bygate, M. (1996). Effects of task repetition: Appraising the developing language of learners. *Challenge and change in language teaching*, 136-146.

Bygate, M. (1999). Task as context for the framing, reframing and unframing of language. *System*, *27*(1), 33-48.

Bygate, M. (2001). Speaking, The Cambridge Guide to Teaching English to Speakers of Other Languages.

CaMLA. (2014). MELAB 2011-2014 Technical Review. Retrieved Online from: http://www.cambridgemichigan.org/wp-content/uploads/2015/07/MELAB-2011-14-technical-review.pdf

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied linguistics*, 1(1), 1-47.

Candlin, C. (1987). Towards task-based language learning. *Language learning tasks*, 5-22.

Carlson, S., & Bridgeman, B. (1986). Testing ESL Student Writers. In Cooper, C. R., & Odell, L. *Evaluating Writing: Describing, Measuring, Judging* (pp. 126-152). Urbana, IL: National Council of Teachers of English.

Cast, B. M. D., (1939). The Efficiency of Different Methods of Marking English Composition. *British Journal of Educational Psychology*, 9, 257-269.

Chaloub-Deville, M. (2003). Fundamentals of ESL admissions tests: MELAB, IELTS, and TOEFL. In D. Douglas (Ed.), *English language testing in US colleges and universities*, (2nd ed., pp. 11-35). Washington DC: NAFSA.

Chapman, M., Collins, C., Dame, B. A., Elliott, H. (2014). A discourse variable approach to measuring prompt effect: Does paired task development lead to comparable writing products? *Research Notes* 55. Cambridge English Language Assessment.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, *19*, 254-272.

Chenoweth, N. A., & Hayes, J. R. (2001). Fluency in writing generating text in L1 and L2. *Written communication*, 18(1), 80-98.

Chiste, K. B. & O'Shea, J. (1988). Patterns of question selection and writing performance of ESL students. *TESOL Quarterly*, 22(4), 681-684.

Connor, U.M. & Carrell, P.L. (1993). The interpretation of tasks by writers and readers in holistically rated direct assessment of writing. In Carson, J.G. & Leki, I. *Reading in the composition classroom: second language perspectives* (pp.141-160). Boston, MA: Heinle & Heinle.

Connor, U. M. & Kramer, M. G. (1995). Writing from sources: Case studies of graduate students in Business Management. In Belcher, D. & Braine, G. *Academic Writing in a Second Language* (pp. 155-182). Norwood, NJ: Ablex.

Connor, U.M., & Lauer, J. (1988). Cross-cultural variation in persuasive student writing. *Writing across languages and cultures: Issues in contrastive rhetoric*, *2*, 138-159.

Cooper, C. R. (1977). Holistic Evaluation of Writing. In Cooper, C. R., & Odell, L. *Evaluating Writing: Describing, Measuring, Judging.* Urbana (pp.3-31). IL: National Council of Teachers of English.

Cooper, C. R., & Odell, L. (1977). *Evaluating Writing: Describing, Measuring, Judging.* Urbana, IL: National Council of Teachers of English.

Council of Europe, (2001). *The Common European Framework of Reference for Languages: learning, teaching, assessment.* Cambridge: Cambridge University Press.

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

Cushing Weigle, S. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.

Davies, M. (2008). *The Corpus of Contemporary American English*: 450 million words, 1990-present. Available online at http://corpus.byu.edu.bnc/.

Diederich, P. B. (1946). The Measurement of Skill in Writing. *The School Review* 10, 584-592.

Diederich, P. B. (1974). *Measuring Growth in English*. Urbana, IL: National Council of Teachers of English.

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). *Factors in Judgments of Writing Ability*. Princeton, NJ: Educational Testing Service.

DiPardo, A. (1994). Stimulated recall in research on writing: An antidote to "I don't know, it was fine. In P. Smagorinsky (Ed) *Speaking about writing: reflections on research methodology* (pp. 162-181). California: Sage Publications Inc.

Eeley, E. G. (1955). The Test Satisfies an Educational Need. *College Board Review* 25, 9-13.

Edgeworth, F. V. (1888). The Statistics of Examinations. *Journal of the Royal Statistical Society* 51(3) 599-635.

Edgeworth, F.V. (1890). The Element of Chance in Competitive Examinations. *Journal of the Royal Statistical Society* 53(3), 460-475.

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

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, 87, 215-251.

Ericsson, K. A., & Simon, H. A. (1985). Protocol Analysis. In T. A. Van Dijk (Ed) *Handbook of discourse analysis*, Vol. 2 Dimensions of Discourse (pp. 259-268) London: Academic Press.

Ericsson, K. A., & Simon, H. A. (1987). Verbal reports on thinking. In C. Faerch & G. Kasper (Eds), *Introspective methods in second-language acquisition research* (pp. 24-53). Clevedon, UK: Multilingual Matters.

Evans, P. (1979). Evaluation of writing in Ontario: Grades 8, 12, and 13. *Review and Evaluation Bulletins,* 1(2). Toronto, Ontario: Canada: The Minister of Education.

Field, A. (2009). *Discovering statistics using SPSS*. Sage publications.

Gabrielson, S., Gordon, B., & Engelhard, Jr., G. (1995). The effect of task choice on the quality of writing obtained in a statewide assessment. *Applied Measurement in Education*, 8(4), 273-290.

Gardner, R. C. (2001). *Psychological statistics using SPSS for Windows*. Prentice Hall.

Garner, R. (1982). Verbal report data on reading strategies. *Journal of Reading Behavior*, 14, 159-167.

Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research.* Mahwah, NJ: Lawrence Erlbaum Associates

Godshalk, F. I., Swineford, F., & Coffman, W. E. (1966). *The Measurement of Writing Ability*. New York: College Entrance Examination Board.

Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing: An applied linguistics perspective*. Harlow, England: Pearson Education.

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

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

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook* (Vol. 5). Cambridge University Press.

Greenberg, K.L. (1981). *The effects of variations in essay questions on the writing of CUNY freshmen*. New York: CUNY Instructional Resource Center.

Greene, S. & Higgins, L. (1994). "Once Upon a Time" The use of retrospective accounts in building a theory of composition. In P. Smagorinsky (Ed) *Speaking about writing: reflections on research methodology*, 8, 115-140.

Hale, G., Taylor, C., Bridgeman, B., Carson, J., Kroll, B., & Kantor, B. (1996). *A study of writing tasks in academic degree programs.* (ETS TOEFL Research Report RR-54). Princeton NJ: Educational Testing Service

Hale, G. A. (1992). *Effects of amount of time allowed on the Test of Written English*. (ETS TOEFL research reports RR-39). Princeton, NJ: Educational Testing Service

Hall, E. (1991). Variations in composing behaviors of academic ESL writers in test and non-test situations. *TESL Canada Journal*, 8, 9-33.

Halliday, M. A. (1985). Dimensions of discourse analysis: grammar. *Handbook of discourse analysis*, *2*, 29-56.

Halliday, M. A., & Hasan, R. (1976). *Cohesion in spoken and written English*. *Longman's, London*.

Hamp-Lyons, L. (1988). The product before: Task-related influences on the writer. In Robinson, P. (Ed.) *Academic writing: process and product* ELT Documents 129. Centre for Applied Language Studies, University of Reading.

Hamp-Lyons, L. (1991a). Basic Concepts. In Hamp-Lyons, L. (Ed.) *Assessing Second Language Writing in Academic Contexts.* Westport, CT; Ablex Publishing.

Hamp-Lyons, L. (1991b).Reconstructing Academic Writing Proficiency. In Hamp-Lyons, L. (Ed.) *Assessing Second Language Writing in Academic Contexts.* Westport, CT; Ablex Publishing.

Hamp-Lyons, L. (1991c). Scoring Procedures for ESL Contexts. In Hamp-Lyons, L. (Ed.) *Assessing Second Language Writing in Academic Contexts.* Westport, CT; Ablex Publishing.

Hamp-Lyons, L. (2002). The scope of writing assessment. *Assessing Writing*, 8, 5-16.

Hamp-Lyons, L., & Mathias, S. P. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing*, 3, 85-96.

Hartog, P. & Rhodes, E.C. (1936). *The Marks and their Examiners*. London: MacMillan

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

Hirokawa, K. & Swales, J. (1986). The effects of modifying the formality level of ESL composition questions. *TESOL Quarterly*, 20(2), 343-345.

Hoetker, J. (1982). Essay Examination Topics and Students' Writing. *College Composition and Communication*, 33(4), 377-392.

Hoetker, J., & Brossell, G. (1989). The effects of systematic variations in essay topics on the writing performance of college freshmen. *College Composition and Communication*, 40(4), 414-421.

Horowitz, D. (1986a). Essay examination prompts and the teaching of academic writing. *English for Specific Purposes*, 5(2), 107-120.

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

Huddleston, E. M. (1954). Measurement of Writing Ability at the College-entrance Level: Objective vs. Subjective Testing Techniques. *Journal of Experimental Education* Volume 22(3), 165-213.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Jennings, M., Fox, J., Graves, B., & Shohamy, E. (1999). The test-takers' choice: an investigation of the effect of topic on language-test performance. *Language Testing* 16(4), 426-456.

Johns, A. M. (1991). Interpreting an English competency examination. The frustrations of an ESL science student. *Written Communication*, 8(3), 379-401.

Johnson, J. & Lim, G. (2009). The influence of rater language background on writing performance assessment. *Language Testing* 26(4), 485-505.

Jung, Y., Crossley, S., & McNamara, D. (2015). Linguistic Features in MELAB Writing Task Performances. CaMLA Working Papers 2015-05. Available Online at: http://www.cambridgemichigan.org/wp-content/uploads/2015/04/CWP-2015-05.pdf

Kandel, I. L. (1936). *Examinations and their Substitutes in the United States*. New York: The Carnegie Foundation for the Advancement of Teaching.

Kincaid, G. L. (1953). *Some factors affecting variations in the quality of students' writing*. Unpublished doctoral dissertation, Michigan State University.

Kroll, B. & Reid, J. (1994). Guidelines for Designing Writing Prompts: Clarifications, Caveats, and Cautions. *Journal of Second Language Writing*, 3(3), 231-255.

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.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied linguistics*, 16(3), 307-322.

Leaper, D. A., & Riazi, M. (2014). The Influence of Prompt on Group Oral Tests. *Language Testing*, *31*(2), 177-204.

Leander, K., & Prior, P. (2004). Speaking and writing: How talk and text interact in situated practices. In C. Bazerman., & P. Prior, *What writing does and how it does it: An introduction to analyzing texts and textual practices* (pp. 201-238). Mahwah, NJ: Lawrence Erlbaum Associates.

Lee, H.K. (2004). A comparative study of ESL writers' performance in a paper-based and a computer-delivered writing test. *Assessing Writing*, 9, 4-26.

Lee, Y. (2002). A comparison of composing processes and written products in timed-essay tests across paper-and-pencil and computer modes. *Assessing Writing*, 8, 135-157.

Lee, Y., Breland, H., Muraki, E. (2004). *Comparability of TOEFL CBT writing prompts for different native language groups* (ETS TOEFL research reports RR-77). Princeton, NJ: Educational Testing Service

Leu, D.J., Keech, K.L., Murphy, S., & Kinzer, C. (1982). Effects of two versions of a writing prompt upon holistic score and writing processes. In J.R. Gray, & L.P. Ruth, *Properties of writing tasks: A study of alternative procedures for holistic writing assessment* (pp. 215-219). Berkeley: University of California, Graduate School of Education, Bay Area Writing Project.

Lim, G. S. (2009). *Prompt and rater effects in second language writing performance assessment* (Doctoral dissertation). University of Michigan.

Lim, G. S. (2010). Investigating prompt effects in writing performance assessment. *Spaan Fellow Working Papers in Second or Foreign Language Assessment.* 8, 95-115.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective* (Vol. 3).

Lloyd-Jones, R. (1977). Primary Trait Scoring. In Cooper, C. R., & Odell, L. *Evaluating Writing: Describing, Measuring, Judging.* Urbana, IL: National Council of Teachers of English.

Lunsford, A.A. (1986). The Past – and Future – of Writing Assessment. In Greenberg, K., Wiener, H. & Donovan R. (Eds.), *Writing Assessment Issues and Strategies* (pp. 1-12*)*. New York: Longman.

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

McNamara, D. S., Louwerse, M. M., & Graesser, A. C. (2002). *Coh-Metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension*. Technical report, Institute for Intelligent Systems, University of Memphis, Memphis, TN.

Mayor, B., Hewings, A., North, S., Swann, J, & Coffin, C. (2007). A linguistic analysis of Chinese and Greek L1 scripts for IELTS Academic Writing Task 2. In Taylor, L. & Falvey, P. (Eds.) *IELTS Collected Papers: Research in Speaking and Writing Assessment* (pp. 250-314). University of Cambridge Local Examinations Syndicate.

Morrow, K. (1979). Communicative language testing: revolution or evolution? In C. J. Brumfit & K. Johnson (Eds.), *The communicative approach to language teaching*. Oxford: Oxford University Press.

Moore, T. & Morton, J. (2005). Dimensions of difference: a comparison of university writing and IELTS writing. *Journal of English for Academic Purposes*, 9 (1), 43-66.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes, *Psychological Review*, 84, 231-259.

O'Loughlin & Wigglesworth, G. (2007). Investigating task design in academic writing prompts. In Taylor, L. & Falvey, P. (Eds.) *IELTS Collected Papers: Research in Speaking and Writing Assessment* (pp. 379-421). University of Cambridge Local Examinations Syndicate.

Ong, J., & Zhang, L. J. (2010). Effects of task complexity on the fluency and lexical complexity in EFL students' argumentative writing. *Journal of Second Language Writing*, *19*(4), 218-233.

Palmer, O. (1961). Sense or Nonsense? The Objective Testing of English Composition. *The English Journal* 50(5), 314-320.

Perelman, L. (2012). Mass market writing assessments as bullshit. In N. Elliott & L. Perelman (Eds.) *Writing assessment in the 21st century* (pp. 425-437). New York: Hampton Press Inc.

Perkins, K. (1980). Using objective methods of attained writing proficiency to discriminate among holistic evaluations. *TESOL quarterly*, 61-69.

Petersen, J. (2009). "This test makes no freaking sense": Criticism, confusion, and frustration in timed writing. *Assessing Writing* 14, 178-193.

Peyton, J. K., Staton, J., Richardson, G., Wolfram, W. (1990). The influence of writing task on ESL students' written production. *Research in the Teaching of English*, 24(2), 142-171.

Plakans, L. (2009). The role of reading strategies in integrated L2 writing tasks. *Journal of English for Academic Purposes*, *8*(4), 252-266.

Polio, C. G. (1997). Measures of linguistic accuracy in second language writing research. *Language learning*, *47*(1), 101-143.

Polio, C. & Glew, M. (1996). ESL writing assessment prompts: How students choose. *Journal of Second Language Writing*, 5, 35-49.

Powers, D. & Fowles, M. (1998). *Test takers' judgments about GRE writing test prompts* (GRE Board Research Report No. 94-13R). Princeton, NJ: Educational Testing Service.

Purpura, J. E. (2005). Michigan English language assessment battery (MELAB). In S. Stoynoff & C. A. Chapelle (Eds.), *ESOL tests and testing* (pp. 87-91). Alexandria, Virginia: TESOL.

Quellmalz, E. S., Capell, F. J., Chou, C. (1982). Effects of discourse and response mode on the measurement of writing competence. *Journal of Educational Measurement*, 19(4), 241-258.

Read, J. (1990). Providing relevant content in an EAP writing test. *English for Specific Purposes*, 9, 109-121.

Robinson, P. (1995). Task complexity and second language narrative discourse. *Language Learning*, 45(1), 99-140.

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

Rose, M. (1980). Rigid rules, inflexible plans, and the stifling of language: A cognitivist analysis of writer's block. *College Composition and Communication*, 31, 389-401.

Rose, M. (1984). *Writer's block: The cognitive dimension*. Carbondale: Southern Illinois University Press.

Ruch, G. (1929). *The Objective or New-Type Examination*. New York: Scott, Foresman and Co.

Ruth, L. & Murphy, S. (1984). Designing Topics for Writing Assessment: Problems of Meaning. *College Composition and Communication*, 35(4), 410-422.

Ruth, L. & Murphy, S. (1988). *Designing Writing Tasks for the Assessment of Writing*. New Jersey: Ablex Publishing Corporation.

Scardamalia, M., & Bereiter, C. (1987). Knowledge telling and knowledge transforming in written composition. *Advances in applied psycholinguistics*, *2*, 142-175.

Shavelson, R. J. (1995). *Statistical reasoning for the behavioral sciences*. Massachusetts: Allyn & Bacon.

Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge: Cambridge University Press.

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied linguistics*, 17(1), 38-62.

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

Skehan, P., & Foster, P. (2001). Cognition and tasks. *Cognition and second language instruction*, 183-205.

Skehan, P., & Foster, P. (2005). Strategic and on-line planning: The influence of surprise information and task time on second language performance. *Planning and task performance in a second language*, 193-216.

Smith, W. L., Hull, G. A., Land Jr. R. E., Moore, M. T., Ball, C., Dunham, D. E., Hickey, L. S., & Ruzich, C. W. (1985). Some effects of varying the structure of a topic on college students' writing. *Written Communication*, 2(1), 73-89.

Spaan, M. (1990). The effect of prompt in essay examinations. In D. Douglas & C. Chapelle (Eds.) *A New Decade of Language Testing Research* (pp. 98-122). Alexandria, VA: TESOL

Stalnaker, J.M. (1934). The Construction and Results of a Twelve-Hour Test in English Composition. *School and Society 39*, 218-224.

Stalnaker, J.M. (1937). Essay Examinations Reliably Read. *School and Society 46*, 671-672.

Stansfield, C. (1986). A history of the Test of Written English: The developmental year. *Language Testing* 3(2), 224-234.

Tavakoli, P. (2009). Investigating task difficulty: learners' and teachers' perceptions. *International Journal of Applied Linguistics* 19(1), 1-25.

Thomson, G.H. & Bailes, S. (1926). The Reliability of Essay Marks. *Forum of Education* 4, 85-91.

Valentine, C. W. (1932). *The Reliability of Examinations*. London: University of London Press.

Way, D. P., Joiner, E. G., **& Seaman, M. A.** (2000). Writing in the secondary foreign language classroom: the effects of prompts and task on novice learners of French: the effects of prompts and task 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), 263-87.

Weigle, S. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, 6(2), 145-178.

Weigle, S. C. (2000). Test review: The Michigan English language assessment battery (MELAB). *Language Testing*, 17(4), 449-455.

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

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Hampshire, UK: Palgrave Macmillan.

Weir, C. J., Vidaković, I., & Galaczi, E. D. (2013). *Measured Constructs: A History of Cambridge English Examinations, 1913-2012* (Vol. 37). Cambridge University Press.

White, E. (1985). Holisticism. *College Composition and Communication* 35, 400-409.

Whithaus, C., Harrison, S. B., & Midyette, J. (2008). Keyboarding compared with handwriting on a high-stakes writing assessment: Student choice of composing medium, raters' perceptions, and text quality. *Assessing Writing* 13, 4-25.

Widdowson, H. G. (1978). *Teaching language as communication.* Oxford: Oxford University Press.

Williamson, M. (1994). The Worship of Efficiency: Untangling Practical and Theoretical Considerations in Writing Assessment*. Assessing Writing* 1, 147-74.

Wiseman, S. (1949). The Marking of English Composition in Grammar School Selection. British Journal of Educational Psychology, 19, 200-209.

Wiseman, S. (1961). *Examinations and English Education*. Manchester, UK: Manchester University Press.

Wiseman, S, & Wrigley, J. (1958). Essay Reliability: The Effect of choice of Essay Title. *Educational and Psychological Measurement*, 18(1), 129-138.

Woodworth, P., & Keech, C. (1980). *The write occasion.* Berkeley: University of California, Graduate School of Education, Bay Area Writing Project.

Yancey, K.B. (1999). Looking Back as we Look Forward: Historicizing Writing Assessment. *College Composition and Communication* 50, 483-503.

Zhang, S. (1987). Cognitive Complexity and Written Production in English as a Second Language. *Language Learning*, 37(4), 469-481.