Fall 11-18-2011

Product and Process in Toefl iBT Independent and Integrated Writing Tasks: A Validation Study

Liang Guo
Georgia State University

PRODUCT AND PROCESS IN TOEFL iBT INDEPENDENT AND INTEGRATED WRITING TASKS: A VALIDATION STUDY

by

Liang Guo

Under the Direction of Sara Weigle

ABSTRACT

This study was conducted to compare the writing performance (writing products and writing processes) of the TOEFL iBT integrated writing task (writing from source texts) with that of the TOEFL iBT independent writing task (writing from prompt only). The study aimed to find out whether writing performance varies with task type, essay scores, and academic experience of test takers, thus clarifying the link between the expected scores and the underlying writing abilities being assessed. The data for the quantitative textual analysis of written products was provided by Educational Testing Service (ETS). The data consisted of scored integrated and independent essays produced by 240 test takers. Coh-Metrix (an automated text analysis tool) was used to analyze the linguistic features of the 480 essays. Statistic analysis results revealed the linguistic features of the essays varied with task type and essay scores. However, the study did not find significant impact of the academic experience of the test takers on most of the linguistic features

investigated. In analyzing the writing process, 20 English as a second language students participated in think-aloud writing sessions. The writing tasks were the same tasks used in the textual analysis section. The writing processes of the 20 participants was coded for individual writing behaviors and compared across the two writing tasks. The writing behaviors identified were also examined in relation to the essay scores and the academic experience of the participants. Results indicated that the writing behaviors varied with task type but not with the essay scores or the academic experience of the participants in general. Therefore, the results of the study provided empirical evidence showing that the two tasks elicited different writing performance, thus justifying the concurrent use of them on a test. Furthermore, the study also validated the scoring rubrics used in evaluating the writing performance and clarified the score meaning. Implications of the current study were also discussed.

INDEX WORDS: Integrated writing task, L2 writing, Writing products, Writing process

PRODUCT AND PROCESS IN TOEFL iBT INDEPENDENT AND INTEGRATED WRITING TASKS: A VALIDATION STUDY

by

LIANG GUO

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2011

PRODUCT AND PROCESS IN TOEFL iBT INDEPENDENT AND INTEGRATED WRITING TASKS: A VALIDATION STUDY

by

LIANG GUO

Committee Chair: Sara Weigle

Committee: Scott Crossley

Viviana Cortes

YouJin Kim

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

December 2011

ACKNOWLEDGEMENTS

This dissertation would never have been completed without the help and support from a great many individuals and a research grant. First and foremost, I am extremely thankful to my committee. My adviser, Dr. Sara Cushing Weigle, introduced me to the world of writing assessment and inspired me to pursue this topic for my dissertation project. The full support and insightful comments I received from her made completing this dissertation project a fruitful and enjoyable journey for me as a researcher and a writer, and the impact is not limited to this project itself. My committee members also provided useful suggestions for my proposal, which greatly improved the quality of this project. A special word of gratitude goes to Dr. Scott Crossley for demystifying Coh-Metrix, the computational tool used in this study, and his willingness and patience to respond to my inquiries regarding statistical analyses. I also owe thanks to Dr. Viviana Cortes and Dr. YouJin Kim for their encouragement and support throughout the dissertation project.

I am also indebted to the teachers in the English as a second language (ESL) Program and the staff at the International Office of Georgia State University, especially Sharon Cavusgil, Sarah Kegley, Cassie Leymarie, Jason Litzenberg, and Audrey Roberson, who helped me considerably for recruiting participants. I would like to extend my gratitude to my colleagues, Man Li, Meg Montee, and Jack Hardy for kindly coding the data and rating the essays. I also need to thank the students who participated in the pilot and final think-aloud writing sessions. Without their willingness to do a great job in these writing sessions, this project would be impossible.

Additional thanks go to my friends Yanbin Lu, Jingsheng Yue, and Julie Konishi for their continued encouragement, support, and friendship. I am also indebted to my parents and my

sister for their unfailing love. Without their support and belief in me, I would not have survived graduate school in the United States.

Finally, I would like to thank Educational Testing Service and The International Research Foundation (TIRF) for providing grants to support the completion of my dissertation project. This project was funded by a TOEFL Small Grant for Doctoral Research in Second or Foreign Language Assessment and TIRF 2011 Doctoral Dissertation Grant.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Writing is considered one of the essential academic skills required in higher education, and its importance also increases as students progress through their years of study (Casanave & Hubbard, 1992). However, measuring writing ability, especially writing ability in a second language (L2), is never an easy task. Considering the role of writing in higher education, the writing ability of L2 writers is very likely to be evaluated in large-scale tests to make decisions as to their preparedness for postsecondary study. In large scale testing situations, independent writing (timed, impromptu-only essay tests) has been widely used as a measure of ESL test takers' academic writing abilities. It is generally agreed that compared with indirect writing assessment (such as multiple choice questions), independent writing tasks provide a more valid representation of underlying writing ability (Camp, 1993). Since essay tests require actual construction of texts, they allow assessment of real writing performance beyond mere analysis and manipulation of morphological and syntactic features of the target language (Camp, 1993; Hamp-Lyons, 1991).

However, concerns have also been raised about writing tests that only contain independent writing tasks (e.g., Cumming, 1997; Hamp-Lyons, 1991; Hamp-Lyons & Kroll, 1996; Lumley, 2005; Weigle, 2004). One of the disadvantages is that independent writing tasks often decontextualize writing activities (Hamp-Lyons & Kroll, 1996). With access only to the prompt, test takers cannot make use of any outside sources beyond their prior knowledge of the imposed topic in text construction (Hamp-Lyons & Kroll, 1996; Wallace, 1997). Therefore, it is argued that independent writing tasks by themselves might not fully reflect real writing activities that are assigned frequently in college study because those writing activities often allow topic selection and involve utilizing background reading support (Braine, 1995; Campbell, 1990;

Carson, 2001; Horowitz, 1991; Kroll, 1979; Weir, 1983). Secondly, writing tests that only contain independent writing tasks are very likely to underrepresent the underlying writing construct as they take a snapshot approach to evaluating writing (Hamp-Lyons & Kroll, 1996; Horowitz, 1991). Constrained by time limits, test takers are not likely to exercise the full range of writing processes including brainstorming, drafting, revising, and editing in one single writing task (Moss, 1994). Thirdly, the use of only one prompt (which is often the case in most writing tests) also casts doubt on the generalizability of writing tests as test takers' writing ability is evaluated based on a single task (or a single genre); therefore, a good sample of the broad range of the underlying writing proficiency cannot be captured (Camp, 1993; Cumming, Kantor, Powers, Santos, & Taylor, 2000; Weigle, 2002).

Given the concerns that have been raised about writing tests that only contain independent writing tasks, integrated writing tasks (using reading and/or listening materials as stimuli for composing) have been proposed as a promising item to be included in writing tests (Feak & Dobson, 1996; Jennings, Fox, Graves, & Shohamy, 1999; Plakans, 2008; Weigle, 2004). For instance, the newer version (Internet-based Test) of Test of English as a Foreign Language (TOEFL iBT) has adopted integrated writing tasks in combination with independent writing tasks in its writing assessment section. The rationale is that the concurrent use of integrated writing tasks and independent writing tasks can enhance the authenticity and validity of ESL writing tests (Cumming, Kantor, Baba, Erdoosy, Eouanzoui, & James, 2005, 2006; Huff, Powers, Kantor, Mollaun, Nissan, & Schedl, 2008).

According to research on academic writing tasks, typical college assignments are unlikely to be completed in isolation (Cumming et al., 2000; Feak & Dobson, 1996; Jennings et al., 1999; Leki & Carson, 1997; Plakans, 2008; Weigle, 2004). Instead, academic writing tends to be

dependent on outside sources. Academic writers are often involved in "text responsible" composing procedure: writing is either based on or stimulated by sources (Carson, 2001). Integrated writing tasks, therefore, not only more authentically resemble the type of writing that is integral to academic contexts of higher education but also better represent the interdependent relationship between reading and writing in academic situations (Cumming et al., 2000; Cumming et al., 2005, 2006; Lewkowicz, 1997; Weigle, 2004). The connection between test performance and targeted language use (academic writing activities) is greatly enhanced by including integrated writing tasks as a task type. With inter-textual activities that connect stimulus materials and the text that test takers construct, the integrated task also provides a more meaningful context similar to real language use in academic settings (Jennings et al., 1999). By better contextualizing the writing activity and better simulating the real academic language use, integrated writing tasks provide a more accurate representation of the real tasks in the target domain, thus building a stronger authenticity argument.

In terms of testing validity, first of all, combined use of integrated writing tasks and independent writing tasks can diversify and improve the measure of writing ability because no single task can be solely reliable to predict the writing ability of a test taker (Cumming et al., 2005; White, 1994). Using the two tasks in combination rather than the independent task or the integrated task by itself, writing tests can obtain a broadened representation of the domain of academic writing (Huff et al., 2008). Different writing tasks tend to involve application of different linguistic abilities because they involve different ways of organizing and conveying information (Cumming et al., 2000; Camp, 1993). On one hand, integrated writing tasks require test takers to respond to source text(s) presented in oral or written format. Test takers are expected to identify and extract relevant information in the source text(s) and organize and

synthesize information (or understanding of this information) in the text they construct (Cumming et al., 2000; Feak & Dobson, 1996). On the other hand, independent writing tasks require an extended written argument built exclusively on test takers' prior knowledge and/or experience. The two tasks are, therefore, expected to be different in the nature of resultant essays.

From an information processing perspective, integrated writing tasks should be different from independent writing tasks in terms of cognitive demands, thus diversifying the measure of writing. With the source material(s) being provided, integrated writing tasks may reduce the cognitive load of searching for content (Hale, Taylor, Bridgeman, Carson, Kroll, & Kantor, 1996). If viewing the two tasks from a knowledge telling or knowledge transforming point of view (Bereiter & Scardamalia, 1985, 1987), the opposite can also be argued. In knowledge telling, writers tend to be familiar with the task and mainly utilize their readily available knowledge (both the content knowledge and the rhetorical knowledge) to address the task. In knowledge transforming, writers are actually using writing to construct new knowledge while responding to the task. Because independent writing is more conventional than integrated writing, test takers tend to be familiar with independent writing and often have the corresponding schema knowledge to respond to the task. In addition, independent writing mainly relies on retrieval of test takers' prior knowledge and/or experience. Therefore, independent writing presumably elicits a knowledge telling writing process and can be regarded as less demanding cognitively. Although in integrated writing, test takers are not really creating new knowledge, they are dealing with newly learned knowledge extracted from source material(s). Due to the unfamiliar content and discourse format (as integrated writing is a newly introduced task type), integrated writing tasks tend to generate more of a knowledge transforming writing process and can be more taxing on test takers' cognitive load. As can be seen, two sides can be argued as to the

comparative cognitive demands of integrated and independent writing. Empirically determining which task is more cognitively demanding is difficult. However, even without knowing the direction of the difference, it is reasonable to conclude that the two tasks have different cognitive demands imposed on test takers.

In terms of possible biases of writing tasks, integrated writing tasks, by providing source materials, might mitigate that negative effect imposed on test takers. With a given topic that is previously unknown, some test takers might lack particular topic knowledge that helps them to successfully complete the independent writing task and, therefore, can be disadvantaged. With background information provided in the source text(s), test takers who lack such knowledge can be better prepared for the writing task (Reid, 1990; Wallace, 1997; Weigle, 2004; Weir, 1983). Many studies have confirmed that compared to independent writing tasks where background knowledge is not provided through stimulus materials, integrated writing tasks are less likely to disadvantage those test takers who might lack related knowledge or experience on the imposed topic (e.g., Jennings et al., 1999; Lee & Anderson, 2007). These studies reinforce the idea that the background information presented in the source materials can help to diminish the unfamiliarity with the assigned topic for the test takers who do not have related topical knowledge.

Studies on the impact of integrated writing tasks have also illustrated that implementing such tasks in assessment can improve the washback on teaching and learning of writing (Esmaeili, 2002; Feak & Dobson, 1996; Weigle, 2004). As previously mentioned, integrated writing tasks better represent the literacy tasks that ESL test takers will face in real academic context. If such writing tasks are included in high stakes exams, teachers and learners are more likely to realize a need in training for skills that relate more to language use in real academic

writing than the formulaic five paragraph writing strategies (Weigle, 2004). Survey studies also revealed that integrated tasks are well accepted by different stakeholders including teachers, test takers, and test users. Such tests are often perceived to be of good task representativeness; they are challenging but reasonable as they match the kind of writing tasks required in academic work (Enright, Bridgeman, Eignor, Kantor, Mollaun, Nissan, Powers, & Schedl, 2008; Feak & Dobson, 1996).

Integrated Writing Tasks

Given the benefits of integrated writing tasks, many exams have utilized source materials (reading and/or listening materials) to stimulate writing. Actually, there are different types of integrated writing tasks that have been put into use. Jamieson, Eignor, Grabe, and Kunnan (2008) divided integrated writing into text-based and situation-based integrated writing. Text-based integrated writing tasks entail construction of a text that summarizes or compares/contrasts information expressed in source materials. The writing is solely based on the information presented in the source materials. The integrated writing task in TOEFL iBT is an example of this type of text-based integrated writing tasks. In situation-based integrated writing tasks, test takers are required to compose emails or letters based on conversations and/or notes communicated either in written or in oral format. An example would be Part One in the writing section of *Certificate in Advanced English* (Jamieson et al., 2008).

In addition to these two types of integrated writing, another integrated writing task that is often used in L2 writing assessment is thematically-related integrated writing. In thematically-related integrated writing tasks, the source material(s) presented and the subsequent writing task are on the same or related topic. Test takers are required to use their own ideas on the topic together with those expressed in the source material(s) while constructing the response essay.

The Georgia State Test of English Proficiency (GSTEP) and the Undergraduate Academic Writing Assessment at the University of Michigan (UAWA) are two of the many examples of thematically-related integrated writing tests (Feak & Dobson, 1996; Weigle, 2004). Surveying exams of L2 academic writing, it can be found that text-based and thematically-related integrated writing tasks are used more often than situation-based integrated writing tasks. One possible reason is that the latter bears less relevance to academic writing assignments than the other two task types.

Purpose of the Study

Despite the many advantages of integrated writing tasks, there have been relatively few studies on these tasks in the literature of L2 writing assessment, especially when compared with the abundance of research on independent writing tasks. Given that other modalities of communication (such as reading and/or listening) are involved in integrated writing, questions have been raised about what such tasks really tap into and whether use of such tasks increases risks of confusing assessment of comprehension with assessment of writing ability (Charge & Taylor, 1997). Additionally, because of the availability of source text(s), validity of integrated writing tasks has also been questioned for the potential verbatim source use, direct language borrowing from source text(s) (Lewkowicz, 1994). Among the few attempts to validate integrated writing tasks, the majority of them have compared such tasks with independent writing tasks. These studies have focused on linguistic features of elicited essays (Cumming et al., 2005; Lewkowicz, 1994), scores assigned (Brown, Hilgers, & Marsella, 1991; Delaney, 2008; Esmaeili, 2002; Lewkowicz, 1994), rater reliability (Weigle, 2004), topic effect (Esmaeili, 2002; Jennings et al., 1999; Lee & Anderson, 2007), use of source text (Cumming et al., 2005; Lewkowicz, 1994), or writing strategies and processes (Esmaeili, 2002; Plakans, 2008).

Although the advantages of integrated writing tasks are often affirmed, it is worth noting that the majority of the related research has mainly looked at thematically-related integrated writing tasks while little is known about the other type of integrated writing tasks that is also integral to academic writing: text-based writing tasks.

Considering the impact of TOEFL in language learning and language instruction in both ESL and English as a foreign language (EFL) contexts, research on its test items is greatly needed. This study thus aims to investigate validity issues of the TOEFL iBT text-based integrated writing tasks. To be specific, this study focuses on exploring the test performance (both linguistic performance and cognitive operations) elicited by the text-based writing tasks, especially in comparison with that in the more traditional independent writing tasks.

Chapelle, Enright, and Jamieson (2008), in building a validity argument for interpretation and uses of TOEFL test scores, draw attention to the link between observed scores and the underlying academic writing abilities. They specify that in order to strengthen the link, evidence related to the discourse characteristics of response essays and to the strategies used to respond to test tasks has to be collected. More specifically, they point out how test performance varies with task types, test scores, and test takers' characteristics and whether it varies in accordance with theoretical expectations are of great importance. First of all, the rationale for the concurrent use of the integrated and independent writing tasks is that the two task types would elicit different writing performance. However, this argument is theory driven. Whether this statement holds still remains unclear and needs empirical data to verify. Secondly, one proposition that underlies the proposed test score interpretation and uses is that academic writing proficiency includes writing products and writing processes test takers use to respond to the writing tasks (Educational Testing Service, 2008). If this proposition holds true, the linguistic features of the resultant

written products and the writing processes that test takers generate are expected to vary with score levels. Again, due to scarcity of research, little is known about whether and how linguistic features and writing processes vary with score levels within text-based integrated writing and how they compare with those of independent writing. Thirdly, if writing tasks tap into academic writing ability, test performance (linguistic knowledge and cognitive operations) are expected to vary along with test takers' exposure to and practice of the target language use. If this is true, it is reasonable to speculate that test takers with more academic experience at the tertiary level of education should outperform those with no or less such experience. This statement should apply even more to the integrated writing if such tasks are better reflective of academic writing tasks assigned in English medium institutions of higher education.

Although all of these speculations are crucial to clarify the link between scores and the underlying construct, which is essential in building a strong validity argument for the score interpretation and uses (Chapelle et al., 2008), they still need empirical evidence to substantiate. As pointed out earlier, text-based integrated writing is underresearched in L2 writing assessment. Therefore, the study aims to explore whether test performance (both the linguistic knowledge and cognitive operations) varies with task type, score levels, and academic experience of test takers in accordance with theoretical expectations, thus building a validity argument for the TOEFL iBT writing tasks.

Context of the Study

Since the study focuses on the TOEFL iBT writing tasks, it is necessary to give a brief introduction of how the integrated and the independent writing tasks are presented in the test. The following information is also available from the official website of TOEFL at http://www.ets.org/toefl.

The TOEFL iBT writing section comprises two writing tasks—writing with/without source text(s). Test takers encounter the text-based integrated writing task first. Test takers are first presented with a short reading passage about 230-300 words long. Three minutes are given to read and comprehend the passage. Then test takers listen to a lecture/conversation, which addresses the same topic but offers a different perspective from the reading material. The listening section usually takes about two minutes. Test takers can take notes during both the reading section and the listening section if they choose to. The integrated writing section (20 minutes long) elicits a compare and contrast essay to summarize how the viewpoints presented in the listening passage relate to those in the reading passage. The typical essays should contain no fewer than 225 words. While composing, test takers have access to the reading passage and their notes. The essay is evaluated holistically on its organization, appropriate and precise use of grammar and vocabulary, and completeness and accuracy of the content covered in the source materials.

Following the integrated writing task is the independent writing task (writing solely based on the writer's own prior knowledge and experiences). Test takers are expected to compose an argumentative essay where they support an opinion on a given topic in 30 minutes. The essay constructed should contain no fewer than 300 words. Test takers are made aware that their writing is graded holistically based on the development, organization, and appropriate and precise use of grammar and vocabulary of their writing. The scoring rubrics for the TOEFL iBT integrated and independent writing tasks are presented in Appendix A.

Research Questions

In order to address the issues of whether test performance varies with task type, essay scores, and academic experience of test takers in the TOEFL iBT writing section, two sets of

research questions are proposed. The first set of questions focus on textual analysis of the writing products of integrated and independent writing. The second set of research questions are mainly about the writing behaviors that test takers go through when constructing their texts in response to the two tasks.

The research questions that guide the quantitative textual analysis section are the following:

1) What linguistic differences and similarities exist in the essays generated in response to the independent writing task and those generated in response to the integrated writing task?
2) Can linguistic features predict essay scores within each task type? If so, what are these features? Are these features different or similar to each other across the two tasks?
3) Does the tertiary level academic experience of test takers have an impact on the linguistic features of the essays they produce in each task? If so, does it have a similar or different impact across the two tasks?

The research questions for the qualitative writing process analysis of the study are stated below:

4) What differences and similarities exist in test takers' writing behaviors when responding to the independent writing task and the integrated writing task?
5) Do writing behaviors employed by test takers vary with essay scores within each task? If so, do they vary in a similar or a different way across the two tasks?

6) Does the tertiary level academic experience of test takers have an impact on the writing behaviors in each task? If so, does it have a similar or different impact across the two tasks?

Significance of the Study

Given that the TESOL iBT often plays a critical role in determining ESL test takers' admission and placement in college study, understanding of the integrated writing task in comparison with the independent writing task is fundamental to the design, development, and use of the test. Through investigating writing products and writing processes within and across the integrated and the independent writing tasks, the study serves three purposes.

First of all, it helps to clarify the construct inherent in the TOEFL iBT text-based integrated writing task by providing both quantitative and qualitative data about the test performance. The results yielded can not only help to clarify the link between observed scores and the underlying writing ability for the TOEFL iBT (Chapelle et al., 2008) but also serve as empirical evidence to substantiate previous theoretical claims made about integrated writing tasks (e.g., its strengthened authenticity). Such information can, therefore, help stakeholders of the test to avoid misconceptions and develop reasonable expectations for independent and text-based integrated writing tasks.

Secondly, by comparing the products and processes of the two tasks, the study can yield systematic evidence to justify the value of combined use of them on the TOEFL iBT. Linguistic and/or process differences to be identified across the two tasks can shed light on the issue whether the two tasks are assessing two different dimensions of the complex underlying writing ability (Delaney, 2008; Esmaeili, 2002). The evidence, therefore, can help to verify whether there is added psychometric value brought about by the inclusion of the integrated writing task.

Such information is not only significant at the theoretical level but also at the practical level. Knowledge of the test performance differences across the two tasks (if there are) is vital to the practical question of whether and why we need to use the independent and the text-based integrated writing task concurrently since two tasks certainly demand more time and resources invested by either the test takers or the test scorers.

Thirdly, by exploring test performance in relation to essay scores and academic experience of test takers, the study helps to clarify score meaning in the integrated and in the independent tasks. This information is also needed to validate the scoring rubrics used in the TOEFL iBT writing section and further illustrate whether and how the two tasks differ when their resultant writing is being evaluated.

To recapitulate, this study contributes to the L2 writing assessment literature by providing systematic evidence about text-based integrated writing especially when compared with independent writing. By combing quantitative and qualitative methods, this study produces more comprehensive descriptive evidence to uncover the construct inherent in the two tasks and to validate the integrated writing task. Detailed information about the product and the process can also make the use of the two tasks more interpretable to writing instructors, test takers, admission faculty and staff, and test designers.

CHAPTER 2

LITERATURE REVIEW

Despite a widespread belief in the value of adding integrated tasks in writing assessment as mentioned earlier, the introduction of such tasks does not come without a host of challenges regarding the skills they tap into, especially when compared to independent writing tasks. Studies thus have been undertaken to examine and validate integrated writing tasks. This chapter reviews pertinent literature regarding validation of integrated writing tasks with or without comparing them to independent writing tasks. Validation efforts have been made in three lines of research, which are discussed accordingly in the three sections of this chapter. The first section focuses on the validation studies that have primarily looked at the scores assigned on integrated essays and how they relate to test takers' independent writing scores, reading scores, and general language proficiency. The following section details studies that have taken a textual approach to validate integrated writing tasks. Studies that have compared the textual features of the essays test takers produce in integrated writing tasks with those in the independent writing tasks are reviewed. Furthermore, studies that have looked at how the textual features relate to scores assigned in integrated writing tasks are also discussed. The third section reviews studies that have investigated the writing processes elicited by integrated writing tasks. This chapter also includes a review of the computational tool—Coh-Metrix (McNamara, Ozuru, Graesser, & Louwerse, 2006) that was utilized in analyzing the textual features of the essays and a review of think-aloud protocols (TAPs) used to collect the qualitative writing process data in the study.

Validation Studies through Score Analysis

To validate and substantiate the claims about integrated writing tasks, only a few studies have been undertaken to explore how integrated writing performance is related to writers'

reading performance, independent writing performance, L2 proficiency, and educational level. Given that different integrated writing tasks are available, text-based integrated writing scores have also been compared with thematically-related integrated writing scores to clarify the underlying construct being measured.

Integrated Writing Scores and Independent Writing Scores

In order to validate integrated writing tasks, particularly in justifying their value when being used together with independent writing tasks on the test, some studies have focused on direct comparisons of scores assigned on the two types of writing tasks. In general, no agreement on the relationship between the scores assigned in the two tasks has been achieved. For example, Lewkowicz (1994) compared the holistic scores that a group of English as a foreign language (EFL) students received on a thematically-related integrated writing task and on an independent writing task when the same scoring rubric was used. Lewkowicz reported no significant difference in the scores. Using two different scoring rubrics, Gebril (2006), also found a high correlation between the two sets of scores when comparing the performance of a group of EFL students on a thematically-related integrated task and an independent writing task.

However, opposite results have also been reported in the literature relating to thematically-related integrated and traditional independent writing tasks. In Delaney (2008), English as a second language (ESL), EFL, and English native speaking writers were required to complete a battery of writing tests including a thematically-related integrated test, a text-based integrated writing test, an independent test, and a reading test. With separate scoring rubrics, the integrated writing performance and the independent writing performance were both assessed by experienced raters. By performing Pearson coefficient analysis on the scores, Delaney found that the independent scores were not significantly correlated with the thematically-related writing test

(r = .12) or with the text-based integrated writing test (r = .20). Similarly, Esmaeili (2002) also compared thematically-related integrated writing scores with independent writing scores and found that the ESL participants achieved significantly higher scores in the integrated than in the independent writing task (F = 134.28, p = .001). The results suggested that when there is a thematic link between reading and writing activities, the writing scores improved significantly as compared to those in activities without a thematic link.

Only limited research can be found on integrated writing tasks for their score relationship with independent writing tasks, and the results yielded do not seem to be in accordance. However, it is worth pointing out that the different research designs (within- or between-subject), scoring systems, and statistical tools used might make direct comparisons between these studies problematic.

Integrated Writing Scores and Reading Scores

Although listening can be part of integrated writing tasks, the majority of integrated writing tasks that have been investigated are reading-to-write activities. Given that reading is actively involved in the performance elicited by integrated reading-to-write tasks, in order to clarify the construct being assessed in such tasks, the scores assigned have also been explored for their relationship with test takers' reading proficiency. Existent research has once again yielded inconclusive evidence about the relationship. Large correlations have been identified in Trites and McGroarty (2005; r =.69 where reading scores were derived from Nelson-Denny task) and in Enright, Bridgeman, and Cline (2002; r =.80 where the reading scores were TOEFL reading scores). Other studies, however, report that although there is a strong relationship between reading and integrated writing scores, the reading score on its own cannot fully capture the integrated writing task scores including both thematically-related and text-based integrated tasks.

For instance, Delaney (2008) reported that although reading proficiency of the participants was significantly correlated with the integrated writing scores (text-based integrated writing and reading, r = .28; thematically-related integrated writing and reading, r = .38), the correlations were weak.

Integrated Writing Scores and General Language Proficiency

The relationship between general language proficiency and integrated writing performance has also been explored. For example, Delaney (2008) also investigated how integrated writing task scores correlated with test takers' general language proficiency. As mentioned earlier, both native and non-native speakers participated in the study. Based on their TOEFL scores, the non-native speakers were further divided into advanced and intermediate proficiency groups. It was confirmed that general language proficiency has a positive correlation with integrated writing performance as the native speakers had the highest mean score among all the groups in both the thematically-related integrated essay writing while the intermediate students had the lowest. However, this significant impact of language proficiency was only valid for the thematically-related integrated writing task but not for the text-based integrated writing task.

Integrated Writing Scores and Educational Levels

As students with more academic experience are supposed to be more familiar with integrated writing tasks, the effect of educational level on performance has been also examined in integrated writing tasks. The general finding is that the educational level, if operationalized as graduate and undergraduate students, produced a significant effect on the integrated writing task scores with the graduate students outperforming the undergraduate students (Delaney, 2008; Trites & McGroarty, 2005). In Delaney (2008), it was further clarified that the difference was

only significant in the thematically-related but not for the text-based integrated writing tasks. The reason speculated is that compared with text-based integrated writing, thematically-related integrated tasks are cognitively more demanding and would require more linguistic resources and more academic experience to enable reader/writers to express and structure content that satisfies the task requirements. Plakans (2010), from a task representation perspective, actually confirmed that her ESL participants' experience with integrated academic writing impacted how they interpreted the task requirement and how they interacted with the source material(s).

Thematically-related and Text-based Integrated Writing Scores

To the researcher's knowledge, there is only one study that has specifically compared performance in different types of integrated writing tasks (thematically-related and text-based). Delaney (2008) correlated text-based integrated writing task scores with those in a thematically-related integrated task. A coefficient of $r = .38$ ($p < .05$) was found between the two sets of scores, indicating that the variance in the two measures overlapped by 14.4%. It was thus argued that performance on the text-based integrated task and on the thematically-related integrated tasks could be considered as two different dimensions of integrated writing ability.

Summary

Through reviewing, it can be found that only a small number of studies have been undertaken to study integrated writing scores. Within the very few investigations, the majority have focused on thematically-related integrated writing tasks while little is known about text-based integrated writing tasks. Furthermore, the findings yielded in the limited studies are often inconclusive with regards to the relationship between integrated writing performance and independent writing performance and reading proficiency. As for the influence of general language proficiency and educational level on integrated writing performance, significant impact

has sometimes been identified for thematically-related but not for text-based integrated writing tasks.

Validation Studies through Textual Analysis

In L2 writing assessment, many efforts have been made to investigate the influence of task type on textual features of resultant writing products at the lexical, syntactic, and discourse levels. The majority of the research done in this area has focused on the effect of the imposed genre (a letter, an essay, or a lab report) and the discourse mode (e.g., descriptive, expository, or argumentative) on the textual features of the essays generated. To the researcher's knowledge, little systematic evidence has been available to answer if and in what ways the features of writing that test takers produce for integrated tasks differ from those they write for independent tasks. Among the very few studies that did investigate the writing products of integrated and independent writing tasks, analysis and comparison has been conducted at levels of text length, lexical sophistication, syntactic complexity, grammatical accuracy, rhetorical and discoursal features, and integration of source materials.

Text Length

Only two studies have been located that have examined whether text length varies systematically across task type (independent vs. integrated), and they yielded inconclusive evidence. In a study of TOEFL prototype tasks, Cumming et al. (2005, 2006) explored three types of tasks: independent and thematically-related reading-to-write and listening-to-write tasks. The researchers found significant difference in text length (defined as total number of words) across the three task types with the independent writing task generating significantly longer texts. Lewkowicz (1994) also investigated text length between a thematically-related integrated task and an independent writing task but found an opposite result from Cumming et al. (2005, 2006).

Given that different time duration was assigned for the tasks, the evidence does not directly lead to a straightforward explanation of the correlation between task type and the word count of the corresponding written outcome as text length may have been limited by the time restriction.

Lexical Sophistication

Analyses have also been untaken to compare the two task types (integrated writing tasks and independent writing tasks) at the lexical level. In Cumming et al. (2005, 2006), lexical sophistication was defined as average word length and type/token ratio (TTR). It was found that both indicators demonstrated statistically significant differences across the task types with both integrated writing tasks generating longer words and higher TTRs than the independent writing task. The explanation provided by the researchers was that in integrated writing, the test takers were more likely to employ words from the source texts directly and the specific topics of integrated writing might inherently involve repetition of certain words.

Syntactic Complexity

Cumming et al. (2005, 2006) investigated syntactic complexity through average number of words per T-unit and number of clauses per T-unit. For number of words per T-unit, Cumming et al. found that this particular feature only differs significantly between the independent writing and the listening-to-write tasks but not between the independent writing and the reading-to-write tasks. For number of clauses per T-unit, significant differences only occurred between the independent writing and reading-to-write tasks but not between the independent writing and the listening-to-write tasks.

Grammatical Accuracy

In Cumming et al. (2005, 2006), grammatical accuracy of the written products was also analyzed for writing task effect. A separate score (on a scale of 3 points) for grammatical

accuracy was assigned by experienced raters to the written products generated under each of the three task contexts (independent, reading-to-write, and listening-to-write). A non-parametric form of multivariate analysis of variance did not show a significant effect on grammatical accuracy for task type.

Rhetorical and Discourse Features

Other researchers when comparing writing features in integrated writing and in independent writing tasks have also looked at more sophisticated discourse features such as idea selection and development in the written products. Lewkowicz (1994) found that there was a significant difference in the number of linguistic propositions used to support arguments in the essays produced in the two task contexts. The researcher further specified that although outside source provided the test takers with ideas, it did not necessarily improve the overall quality of the argument. In the integrated writing task, the test takers tended not to fully elaborate the ideas extracted from the source material. The finding is in line with Campbell (1990) that the test takers often used ideas from source texts as new propositions rather than as support to their own ideas. Along the same lines, Watanabe (2001) reported that essays generated in response to the thematically-related integrated writing tasks often contained fewer original theses when compared with those in response to the independent writing tasks.

Integration of Source Materials

A few research studies have also examined whether test takers specified the source of the incorporated information in the integrated writing tasks. For example, Cumming et al. (2005, 2006) compared the source citing across reading-to-write and listening-to-write tasks. As for how test takers integrate source materials into their own writing, a few researchers have looked at L2 writers' integration style—whether the content of source information was presented in

declaration, quotation, paraphrase, or summary. The general pattern that emerged is that for the independent writing task, the test takers mainly used declaration for expressing the message from the prompt but in the integrated writing tasks, the test takers tended to rely on paraphrase and summary to integrate information from source text(s) (Cumming et al., 2005 , 2006; Watanabe, 2001).

Textual Features in Relation to Essay Scores

The previous section discusses the studies that have compared linguistic differences between integrated and independent writing. Meanwhile, there are also studies that focus only on linguistic features within integrated writing tasks. These studies often related linguistic features to writing proficiency (as reflected in the holistic scores assigned) to validate integrated tasks. Whether and how the textual features vary in the essays produced by test takers at different proficiency levels is of great importance, as this information is needed to verify the role of linguistic features of integrated writing in characterizing L2 writing proficiency and to validate the scoring schemes being used to assess test takers' performance on integrated tasks (Cumming et al., 2005). Text length, lexical sophistication, syntactic complexity, grammatical accuracy, and integration of source materials have all been examined to see whether they vary with writing proficiency. In order to illustrate possible similarities and differences between integrated and independent writing tasks, studies that have related textual features to writing proficiency in independent writing tasks are also reviewed.

Text Length

Gebril and Plakans (2009) and Watanabe (2001) looked at text length in relation to the holistic scores assigned on thematically-related integrated essays. Both studies demonstrated that text length has a significant effect on the score levels; thus they concluded that the longer essays

were evaluated more favorably by raters in integrated writing tasks. Similar findings have been repeatedly reported with independent writing tasks, indicating a strong and direct correlation between text length and human judgments of independent writing quality (e.g., Carlson, Bridgeman, Camp, & Waanders, 1985; Frase, Faletti, Ginther, & Grant, 1999; Grant & Ginther, 2000; Reid, 1986).

Lexical Sophistication

Cumming et al. (2005, 2006) investigated lexical sophistication through TTR and average word length. In relation to general writing proficiency of the test takers (general writing proficiency being determined by the holistic scores assigned on the independent writing task), the researchers reported that higher proficiency seemed to correlate with higher type/token ratio but not with average word length. However, since the integrated essays were analyzed together with the independent essays, whether the identified relationship between lexical features and writing proficiency will still apply to the integrated essays by themselves is not clear.

Gebril and Plakans (2009) focused only on integrated writing and investigated the effect of integrated writing proficiency (integrated essay scores) on lexical sophistication. Average word length was also used to measure lexical sophistication, and a similar finding was reported, that is when exploring the feature in relation to integrated essay scores, there are no significant differences demonstrated.

Compared to research on integrated writing, L2 independent writing research has explored many more features of lexical sophistication and their effect on the independent essay scores. These features include average word length (Frase et al., 1999), lexical diversity (Crossley & McNamara, in press a; Grant & Ginther, 2000; Reppen, 1994), specific lexical categories (Ferris, 1993), and nominalizations (Connor, 1990). The general tendency identified is

that higher rated independent essays are associated with longer words, greater lexical diversity, more frequent use of nominalizations and certain lexical categories. Even though only very little evidence is available about integrated writing, the research seems to suggest that lexical sophistication is an important predictor of essay scores in both independent and integrated writing.

Syntactic Complexity

Cumming et al. (2005, 2006) analyzed syntactic complexity through two indicators: the number of words per T-unit and the number of clauses per T-unit. As for the first indicator (number of words per T-unit), Cumming et al. reported that there is a main effect for proficiency levels (measured by the holistic scores assigned on the independent essays) when independent and integrated essays were considered at the same time. The proficiency levels did not have a significant effect on the second indicator—the number of clauses per T-unit.

Gebril and Plakans (2009) also explored the interaction between the number of words per T-unit with writing proficiency levels (defined as writing scores assigned on the integrated writing task) within integrated writing itself. In contrast to Cumming et al. (2005, 2006), they, found a slightly different picture with no significant difference being identified. While looking at the mean number of clauses per sentence in relation to the writing proficiency, again, no statistically significant syntactic differences could be found.

Likewise, in independent writing tasks, inconsistent results have also been reported about the influence of syntactic complexity on holistic scores assigned. For example, variety of syntactic patterns (Ferris, 1993) and mean number of words before the main verb (McNamara, Crossley, & McCarthy, 2010) were found to be significant predicators of raters' judgment of ESL writers' writing quality. In Song (2007), however, only a non-significant correlation was

demonstrated between syntactic complexity (measured by means of dependent clauses per clause and clauses per T-unit) and the holistic scores assigned.

Grammatical Accuracy

Grammatical accuracy has also been examined to explore how test takers' essay features relate to writing proficiency levels. As described previously, Cumming et al. (2005, 2006) specifically assigned a score for grammatical accuracy on each essay, either independent or integrated. The studies demonstrated that in relation to the proficiency levels (holistic scores for the independent writing task), there were significant differences in grammatical accuracy scores when independent and integrated essays were analyzed together.

Using the same measurement of grammatical accuracy, Gebril and Plakans (2009) also analyzed whether there were significant differences in grammatical accuracy across different proficiency levels. The study yielded a significant difference and demonstrated that the mean ratings of grammatical accuracy increased with the proficiency level of the test takers.

In independent writing tasks, however, mixed findings have been reported about grammatical accuracy in relation to raters' judgment of writing quality. For example, in Song (2007), grammatical accuracy (calculated by error free T-units and errors per T-unit) was found to be uncorrelated with writing performance. In contrast, Homburg (1984) found that grammatical accuracy (also measured by error free T-units and errors per T-unit) differentiates writing quality.

Integration of Source Materials

Since integrated writing tasks involve the use of source materials, another line of research has focused on how L2 writers extract information from the source material(s) and how they incorporate such information in the texts they construct. Watanabe (2001) explored use of ideas from source materials in relation to writing proficiency within the integrated writing task.

Watanabe noticed that the group of writers with higher writing proficiency tended to utilize information from source texts and fully exploit the extracted information in their own text construction. In contrast, the lower proficiency groups tended to either ignore the information or use direct textual borrowing as a coping strategy in handling information that was not from their prior knowledge or experience. Similarly, Johns and Mayes (1990) also examined whether idea use was correlated with writing proficiency in integrated writing tasks. What they found is that the less proficient L2 writers were less likely to locate the interrelationship of the ideas presented in the source materials.

L2 writers' verbatim source use in the final written outcome has also been explored in relation to their writing proficiency. Verbatim source use is often defined as strings of three words or more in the test takers' scripts that are directly taken from the source text(s). Using computer programs to tally all the occurrences of such strings, studies have found that for integrated reading-to-write tasks, there is often a negative correlation between the human judgments of the writing quality and verbatim source use in the essays (Campbell, 1990; Cumming et al., 2005 & 2006; Currie, 1998; Gebril & Plakans, 2009). However, when integrated listening-to-write task is considered, a similar negative correlation was not reported (Cumming et al., 2005, 2006). No definitive conclusion has been made as to what leads to such a difference. Possible explanations were given that this task type effect may have resulted from multiple factors including test takers' comprehension of the source text, working memory, time allocation for the writing activity, and/or the characteristics of the source materials themselves.

In exploring the use of source information, previous research has also focused on the integrated style (whether the information was incorporated through declaration, quotation, paraphrase, or summary) and on whether the source of integrated information was specified in

test takers' constructed essays. In consideration of the influence of writing proficiency on integration style, Cumming et al (2005, 2006) found that higher rated essays tended to contain more information extracted from the source materials and use summary frequently to incorporate such information while the lower rated essays contained less information from the source materials and mainly relied on verbatim source use while presenting the information

As for specification of the source of incorporated information, Cumming et al (2005, 2006) found that test takers generally tended not to cite the source and this tendency was relatively consistent regardless of the integrated task types and the proficiency levels. This finding has actually been confirmed in many other studies (e.g., Campbell, 1990; Johns & Mayes, 1990; Watanabe, 2001).

Summary

As can be seen from the studies reviewed, task type difference (integrated vs. independent writing) often manifests itself in textual features of the generated writing products. These features include text length and lexical and syntactic features. When it comes to textual features such as syntactic complexity in relation to writing proficiency within individual task types, mixed results have been reported for both integrated writing and independent writing tasks. With regards to incorporation of source materials, higher proficiency test takers, compared to their lower proficiency counterparts, tended to more frequently integrate and exploit more source information in their integrated writing without significant verbatim use of source text(s).

However, it should be noted that there are only a very limited number of studies available on the textual features of integrated writing tasks (in comparison to those on independent writing tasks), especially those that directly compare integrated writing with independent writing. Many of the studies on integrated writing or on independent writing, even in studying the same features,

have defined and measured the features differently and contained great variability in task requirements and task conditions, which adds to the difficulty in reaching any definitive conclusions about the interaction between textual features and task types or perceived writing quality. Furthermore, the majority of the studies described so far have focused on thematically-related integrated tasks while text-based integrated writing tasks, especially those involving multiple sources presented in different modes, have not attracted much attention. Finally, the linguistic features explored in the existent studies are often surface level features such as lexical diversity and syntactic complexity while little is known about whether and how deep level features that tap into cohesion within texts are used in both integrated and independent writing tasks. Collectively, it can be seen that only very limited evidence exists regarding textual differences between integrated writing and independent writing, and even less information is available on text-based integrated writing. Therefore, there is still a lot to be done to depict the features of text-based integrated writing and compare them with those of independent writing, thus validating the new task and justifying its value on the test of writing proficiency especially when it is used simultaneously with the independent writing task.

Validation Studies through Process Analysis

While the previous studies that focused on written products of integrated writing tasks have contributed to our understanding of the use of such tasks, they do not examine the processes that test takers employ in integrated writing. To investigate such issues, we must turn to writing process research, a body of work that utilizes techniques such as retrospective interviews, questionnaires, and think-aloud protocols to explore writers' meaning making processes and cognitive operations. However, in the literature of L2 writing and L2 writing assessment, the main focus of this work has been on providing a general description of processes involved in

independent writing and how they vary with writing expertise of the writers. Only a few studies have looked specifically at integrated writing. The majority of these studies again have focused on thematically-related integrated writing tasks. For this reason, writing process studies on text-based integrated writing tasks in first language (L1) context and in non-testing writing context are also reviewed in the hope that a more comprehensive picture can be provided.

Process Studies on Integrated Tasks in L2 Context

A number of L2 studies have investigated the writing processes used in integrated writing tasks in testing situations. Different methods have been utilized to investigate the process, including retrospective interviews, think-aloud protocols, and checklists of writing strategies. Before a detailed review of the related studies is presented, it should be noted that the terms "writing processes" and "writing strategies," although they can be differentiated by the purposefulness of the writers (Cohen, 1998), are often used interchangeably in the majority of the studies to be reviewed.

Esmaeili (2002) directly compared writing processes in thematically-related integrated and independent writing tasks. 34 ESL students majoring in engineering participated in his study. After finishing the writing tasks, the participants were required to complete retrospective interviews as well as a written checklist of writing strategies. According to the participants' self reports, they relied extensively on the reading text they had read prior to performing the integrated writing task and constantly evaluated the content presented in the source text and adjusted their writing based on the information and structure of the source text. The researcher, therefore, drew the conclusion that since integrated writing requires incorporation of information from outside sources, the writing process involved in integrated writing tasks is interdependent and intertwined with reading components and thus differs significantly from that in the independent writing task.

Plakans (2008) reported on another comparative study on writing process involved in a thematically-related integrated task and an independent writing task. The participants were 10 ESL students (five graduate and five undergraduate students). Think-aloud protocols and interviews were used to elicit information related to the writing processes. Although it was found that test takers differed from each other qualitatively in terms of the writing processes across the two tasks, the general pattern that emerged is that the integrated writing task generated a more interactive process while the independent writing task required more initial and less online planning. For the integrated task, the episodes illustrated in the think-aloud data revealed more online planning and a more recursive and less linear approach to meaning making and meaning construction during the comprehending and composing processes. Furthermore, in discussion of the influence of the academic experience (graduate vs. undergraduate) on writing processes in the integrated writing, the researcher stated that more experienced writers (graduate student writers) tended to employ a more interactive writing process in completing the integrated writing task. A writing process model was also proposed in the study for the integrated writing tasks. The model includes two major stages: prewriting (reading and planning stage) and writing. For the prewriting state, the writers prepared themselves for the writing task following a linear process of comprehending the task prompt and instructions, analyzing the task, comprehending the source text, and mining information for use in writing. In the second stage (the writing stage), the writers followed a series of non-linear processes to construct and revise their text including planning, rehearsing phrases, rereading source materials, and examining mechanics and language use.

Ascencion (2005) also looked at the writing processes in integrated writing. Instead of comparing and contrasting independent writing processes with integrated writing processes, the

study focused on two different types of integrated writing tasks: text-based integrated writing and thematically-related integrated writing tasks. In addition to an attempt to clarify the construct being assessed in the two types of integrated writing tasks, the study also aimed to show the interaction between general writing proficiency and writing processes in the integrated writing tasks. Six advanced ESL writers and three less experienced EFL writers were asked to use think-aloud protocols to verbally report their writing processes. The researcher applied the categories of planning, monitoring, organizing, selecting, and connecting (developed by Spivey (1984) in his discourse synthesis model) to code the writing process data. The findings from the think-aloud protocols confirmed the model proposed by Spivey. When it came to the frequencies of each category, it was specified that the most frequently used strategy was monitoring followed by planning, organizing, selecting and connecting.

In the discussion of the differences between the two integrated writing tasks (text-based vs. thematically-related tasks), Asceonion (2005) reported that the process data generally confirmed the construct described in the two integrated writing tasks. The process data illustrated that the two tasks did focus on different aspects of the underlying writing construct with the thematically-related integrated task eliciting more cognitive operations than the text-based integrated writing (summary writing). The participants were found to monitor their reading comprehension when they processed the source materials in the thematically-related integrated writing task more closely than in the text-based integrated writing task. They also did more planning on form and content when engaged in the thematically-related integrated writing task than in the text-based summary writing task.

Compared with the EFL group, the experienced ESL group was found to spend more time planning their content and was more involved in interacting with the source text. On the

other hand, in carrying out the integrated writing tasks, the EFL group was found to be more concerned with composing their own text (concentrating on linguistic features in text construction) and was more aware of language related issues and difficulties.

Yang (2009) specifically investigated the writing strategies utilized by test takers in responding to a TOEFL iBT integrated writing task. She asked 161 ESL students to respond to a checklist of writing strategies immediately after they completed the text-based integrated writing task (reading-listening-to-write). Information elicited by the checklist together with the retrospective interview data was used to analyze the test takers' strategy use and its relationship with their performance (the performance was operationalized as their scores) on the TOEFL iBT integrated writing task. The strategies investigated included rhetorical strategies (organizing, selecting, and connecting), self-regulatory strategies (planning, monitoring, and evaluating), and test taking strategies (test management and test wiseness).

After conducting a reliability test of the items on the checklist, test management strategies were excluded from further analysis. In general, the study showed that rhetorical strategy use had a positive direct impact on the writing performance while test-wiseness strategy use had a significant negative effect. Self-regulatory strategy use was shown to have an indirect positive impact on the test takers' writing performance via rhetorical strategy use and an indirect negative influence via the use of test-wiseness strategies.

Strategy use of high performance and low performance groups was also compared in Yang (2009). The study drew attention to the finding that the two groups used similar types of strategies while carrying out the integrated writing task. However, the frequency and quality of the strategy use led to the different writing quality. For example, in the reading phase, the high performance group was found to be engaged more in global reading while the low performance

group struggled more with lexical and sentential decoding. In the listening phase, although both groups reported note-taking behaviors, the quality of notes differed greatly as the low performance group tended to write down unfamiliar words or phrases for later use in text construction as a result of their lack of L2 language proficiency.

Process Studies on Integrated Tasks in L1 and Non-testing Writing Context

As can be gleaned from the previous review on integrated writing in L2 context, there is only a very limited number of studies that have addressed writing processes involved in integrated writing, and the majority of them are about thematically-related integrated writing tasks. Text-based integrated writing is far from being extensively studied. Therefore, research on summary writing in the context of L1 writing and non-testing writing will be reviewed next to shed more light on that particular integrated writing task.

van Dijk and Kintsch (1977) (cited in Kintasch & van Dijk, 1978) proposed one of the most influential models of summary writing by analyzing writing processes used by L1 writers. The model focuses on the procedures that reader/writers go through to move from the source text to the target text, and it specifies that reproducing perceived macrostructure of the source text and generating inferences in construction of the new text are the key in summary writing. To be specific, three processes are involved in summary writing including a) deletion, b) generalization of irrelevant or redundant propositions, and c) integration (constructing new inferred propositions). The researchers emphasized that these processes are used by writers in successful construction of target text-based on the source text.

Based on the model of van Jijk and Kintsch (1977), Brown and Day (1983) put forth a more complex classification of the processes in summary writing. According to them, there are six processes involved in summary writing including: a) deleting trivial information, b) deleting

redundant materials, c) substituting a superordinate (substitution of a category name for instances of a category), d) integrating, e) selecting a topic sentence (near verbatim use of a topic sentence from the source text), and f) inventing (creating and using a topic sentence that was not readily presented in the source text).

Using think-aloud protocols with English native speaking participants, Brown and Day (1983) confirmed that all six processes were utilized in summary writing. Furthermore, they reported that a developmental pattern was identified in the participants' use of the six processes. It was noticed that the writers at all proficiency levels were able to successfully identify trivial information to delete, but the higher proficiency participants tended to outperform their less experienced counterparts in the use of more complex processes such as the use of superordinate substitution and invention.

The previous two studies focused on specification of the processes involved in summary writing. In addition, researchers have also related the occurrences of writing processes to the writing experience of the writers. Kennedy (1985), for instance, compared the writing processes of novice writers and proficient writers in summarizing. In the study, native English speaking college students were asked to read three related articles and write an objective essay based on the given material. Kennedy found that the proficient writers were active readers and note takers. They tended to revise their notes before incorporating them into their own writing. On the other hand, the novice writers read more passively and did not extensively interact with the source text.

Taylor and Beach (1984) reported another comparative study involving, this time, inexperienced and professional writers while performing a summary writing task on an expository source text. It was found that the two groups of participants differed from each other in the reading process as well as in the writing process. In the reading process, the professional

writers were more careful readers and studied the text till they were convinced that they fully understood the text. Prior to writing, they spent more time planning than the inexperienced writers. In composing the summary, the professional writers monitored the source text more constantly to check for accuracy and were more objective in presenting the ideas conveyed in the source text than the inexperienced writers. The professional writers also took audience into consideration when making decisions about the level of generality in their summarized text.

Yang and Shi (2003) looked at text-based integrated writing in non-testing context. Six first-year Master of Business Administration students (three ESL and three native speaking students) participated in think-aloud sessions while completing a course-related summary task. Building on the Hayes & Flower (1980) model of cognitive processes of writing, the authors proposed a four category model of integrated writing. The four categories are planning, composing, editing, and commenting. Within each category, different strategies are also identified. For instance, under the general category of planning, there are subcategories including planning for organization, planning for content, planning for text format, planning for word and sentence choices, and reviewing task requirements. Using this four category coding scheme, the researchers reported that the participants' most frequently used strategies include verbalizing what is being written, planning content, referring to the sources, reading what has been written, reviewing and modifying one's writing, and commenting on the source texts. It was also mentioned that the participants' previous writing expertise in disciplinary writing and their perceptions of the writing task greatly impacted their writing processes and use of writing strategies.

Summary

Although only limited studies can be found on integrated writing in L2 context, the general conclusion that can be drawn is that thematically-related integrated tasks often entail test takers' active interaction with source text(s), especially for expert or more experienced writers. Experienced writers tend to be very engaged with comprehending and incorporating ideas from source materials instead of focusing on decoding at the sentence level. L1 studies that have looked at summary writing (text-based integrated writing) in particular have also been reviewed. Using between-subject research design, the related studies have also illustrated that writing experience and expertise have an effect on the processes that writers employ when responding to a text-based integrated writing task.

As noted earlier in the product studies, text-based integrated writing is also underrepresented in process-oriented research in L2 writing context. Very few studies have looked at text-based integrated writing with L2 writers, especially when there is more than one source text available. Therefore, research is still needed to gain more insight into the process aspect of the underlying construct being assessed in text-based integrated writing tasks.

Summary of Validation Studies

Despite their many potential benefits, there have been relatively few studies of integrated writing tasks in the literature of L2 writing assessment, especially when compared with the abundance of research on independent writing tasks. Among the few studies that have addressed the integrated writing tasks, many of them have attempted to validate integrated writing tasks by examining the integrated scores in relation to other measures of language proficiency (Brown et al., 1991; Delaney, 2008; Esmaeili, 2002; Lewkowicz, 1994), analyzing textual features in relation to task types and essay scores (Cumming et al., 2005 & 2006; Esmaeili, 2002; Gebril &

Plakans, 2009; Lewkowicz, 1994), or investigating the writing processes employed in constructing integrated essays (Esmaeili, 2002; Plakans, 2008; Yang & Shi, 2003).

These studies often affirmed the advantages of including integrated writing tasks in writing assessment such as the involvement of meaningful interaction with the source materials and thus contributed to our understanding of integrated writing. It is, however, worth noting that the review also reflects several gaps with the existent research pertaining to integrated writing tasks. First of all, the majority of the research on integrated writing has looked at thematically-related integrated writing tasks while little is known about the other type of integrated writing tasks: text-based writing tasks. Second, previous studies have focused either on the product or on the process of the writing performance to validate integrated writing tasks used in testing context. Very few studies have incorporated quantitative (product) and qualitative (process) data together to build a more comprehensive picture to validate integrated writing tasks (Bachman & Palmer, 1996; Cumming et al., 2000). Third, in investigating writing products of integrated writing, only surface level features have been reported while little is known about deep level linguistic features that contribute to cohesion within texts. Therefore, validation studies of integrated writing tasks, especially studies of text-based integrated writing, are still needed. To clarify the construct inherent in text-based integrated writing and verify the previous statements that have been made about text-based integrated writing tasks as a promising test item, it is necessary to conduct textual analysis of the essays composed by test takers and obtain qualitative information on the test taking processes as well (Bachman, 2004).

In the following sections, the computation tool used in the quantitative analysis section of the study and the TAP used in the qualitative analysis section will be reviewed.

Coh-Metrix

In order to explore the linguistic features of the computerized integrated and independent essays, an automated textual analysis tool, Coh-Metrix, was used in the study. Before the use of computational tools in examination of L2 writing, hand counts and subjective judgments were often employed in documenting and analyzing linguistic features. The results yielded through such approaches have contributed to our understanding of L2 writing, but these approaches are often very time-consuming, laborious, and prone to mistakes. Computational analysis, although it has its own limitations (Ferris, 1993; Frase et al., 1999), is more efficient and accurate, and the data it generates is more consistent and comprehensive especially when dealing with texts in large quantities (Crossley & McNamara, 2009).

The Biber tagger (Biber, 1988, 1995) and the STYLEFILES (Reid, 1992) are two of the computational tools that have been applied to automated analysis of L2 writing. These computational tools, however, mainly draw on surface measures of linguistic features such as TTR, word length, and perfect aspect verbs. Although studies on these surface features permit insights into the nature of L2 writing, these measures fail to account for more sophisticated linguistic features and deep level textual properties such as cohesion. Unlike these computational tools, Coh-Metrix synthesizes many advances in various disciplines and approaches such as computational linguistics, corpus linguistics, psycholinguistics, and discourse processing (Crossley & McNamara, in press a). To generate a comprehensive evaluation of given texts, Coh-Metrix integrates many devices including lexicons, pattern classifiers, part-of-speech (POS) taggers, syntactic parsers, and shallow semantic interpreters. To illustrate, Coh-Metrix draws on the Medical Research Council (MRC) Psycholinguistic Database (Coltheart, 1981) to report psycholinguistic information about words including word concreteness and word familiarity. It

also utilizes latent semantic analysis (LSA), a mathematical and statistical technique that represents deeper world knowledge based on large corpora of texts (Crossley & McNamara, in press a), to track semantic similarity between words.

For these reasons, Coh-Metrix enjoys many advantages. First of all, the power of Coh-Metrix allows for quantitative examination of surface level linguistic features as well as deep level features related to textual cohesion. Secondly, some of the indices reported by Coh-Metrix (such as syntactic complexity indices and word overlap indices) have not been available in previous computational analysis before (Graesser, McNamara, Louwerse, & Cai, 2004). Thirdly, by adopting the most recent developments in different fields related to linguistics, Coh-Metrix avoids some problems associated with more traditional methods for measuring linguistic features. For instance, TTR is often found to be unreliable because of its heavy reliance on text length. Instead of relying on TTR, Coh-Metrix reports lexical diversity through more reliable and valid indices such as the Measure of Textual Lexical Diversity (MTLD), calculated as the mean length of word strings that maintain a criterion level of lexical variation (McCarthy & Javis, 2010).

To be specific, Coh-Metrix measures many aspects of input texts including basic text information (e.g., number of words, number of paragraphs), lexical sophistication (e.g., word concreteness, imagability, word polysemy values, word frequency, lexical diversity, etc), syntactic complexity (e.g., mean number of words before the main verb, syntactic similarity, etc), and cohesion such as causality and lexical overlap (Crossley & McNamara, in press; Jurafsky & Martin, 2002). Each aspect of the input texts is evaluated through many measures. Even when reporting on the same measure, Coh-Metrix in many cases assesses it with different indices. For instance, causality is assessed by causal verbs, causal connectives, or a combination of the two. That same feature can also be reported on different levels (on text, paragraph, or sentence levels

or for content or all words). In some cases, the boundaries between Coh-Metrix indices are not clear cut but interrelated. For example, greater lexical diversity indicates higher lexical sophistication of a given text because a more diverse range of words are being used. However, at the same time, lexical diversity is also related to textual cohesion because greater lexical diversity signifies less word overlap and thus lower lexical cohesion among sentences of a given text.

In textual analysis, Coh-Metrix has been used with great success to determine a wide range of linguistic differences between and within text types. Several studies, for example, have used Coh-Metrix to identify lexical, cohesive, and/or syntactic differences between different text types including simplified and original texts (Crossley, Louwerse, McCarthy, & McNamara, 2007), texts written by different authors (McCarthy, Lewis, Dufty, & McNamara, 2006), English essays produced by L2 writers from different linguistic backgrounds (Crossley & McNamara, 2009), and student essays at different proficiency levels (Crossley & McNamara, in press a). Within one text type, Coh-Metrix has also been employed to illustrate differences among various sections of the text (Graesser, Jeon, Yang, & Cai, 2007; Lightman, McCarthy, Dufty, & McNamara, 2007). Taken together, these studies demonstrate the effectiveness and efficiency of Coh-Metrix as a computational tool for assessing and differentiating text types not only in L1 writing but in L2 writing as well. However, similar to many other computational tools, there is also limitation with Coh-Metrix in analyzing L2 writing. It does not measure or report language errors which are often associated with L2 writing. Therefore, it is important to be aware that in analyzing L2 writing, there is a potential risk that language errors might not animate computational analysis in some indices.

TAPs

This study employed TAPs to elicit information about the participants' cognitive operations while they were constructing texts in response to the two writing tasks: the independent and the integrated writing tasks. TAPs require participants to keep producing verbal reports of their mental processes without explaining or justifying them (Ericsson & Simons, 1993). In think-aloud writing sessions, participants are expected to verbally report everything that goes through their minds while they are performing the writing tasks.

TAPs, used to explore cognitive operations of writers, help researchers to gain access to rich data about why and how writers respond to a writing task in the way it is: how they interpret a writing task, the decisions they make, and the thoughts that govern these decisions (Faerch & Kasper, 1987; Kormos, 1988; Swarts, Flower, & Hayes, 1984). TAPs have been used extensively in research on cognitive processes involved in writing. Researchers have used TAPs to construct models of writing processes (e.g., Flower & Hayes, 1981), to study task interpretation of test takers (e.g., Connor & Carrell, 1993), and to explain the differences between skilled and novice writers (e.g., Plakans, 2007). The use of TAPs not only provides "direct evidence about processes that are otherwise invisible" (Cohen, 1987, p. 91) but also greatly supplements conventional quantitative approaches adopted in test validation studies (Green, 1998).

Compared with other self-reported methods, such as retrospective checklists and interviews, TAPs have the advantage of being immediate. Retrospective methods, due to the time lag, add to the difficulty for writers to fully retrieve all the cognitive operations and increase the possibility for reporting what they believe they do (Ericson & Simons, 1987; Green, 1998). On the other hand, TAPs record information about cognitive operations in real time, and thus the data yielded is expected to better reflect what writers actually do (Swarts et al., 1984). Therefore,

due to the real time recording, data collected through TAPs illustrate more specific instances of actual behavior rather than participants' or researchers' generalized statements about what individuals are doing (Cohen, 1998; Ericsson & Simon, 1987; Green, 1998).

Although TAPs significantly promote investigation of cognitive processes, the method has its limitations. Verbally reporting cognitive operations might be distracting and unnatural when individuals are focused on completing the given task (Stratman & Hamp-Lyons, 1994). Individuals, especially L2 users, might not be used to verbalizing their internal thoughts (Sasaki, 2000). The distraction and unnaturalness of articulating thoughts has been pointed out to run the risk of veridicality and reactivity. Veridicality refers to whether TAP generated data can truly and completely represent all the mental thoughts that participants experience. In other words, veridicality concerns whether participants report everything that comes to their minds without omission and modification (Ericsson & Simon, 1987; Stratman & Hamp-Lyons, 1994). Reactivity concerns whether the process of verbally reporting alters the process being observed and the outcome it elicits (Ericsson & Simon, 1987; Stratman & Hamp-Lyons, 1994).

Ericsson and Simon (1993) specifically addressed these criticisms. To argue against the threat of veridicality, the researchers offered both theoretical arguments and empirical evidence. First of all, they acknowledged possible incompleteness of TAP data because certain internal thoughts are automatized or related to long-term memory and thus not accessible for verbalization. Using evidence from cognitive psychology, they argued that in problem solving, individuals mainly use short-term memory and data collected through TAPs actually reflect what they explicitly attend to. Therefore, TAP data is a valid and reliable representation of individuals' mental thoughts. Furthermore, while acknowledging that TAP data might be incomplete, Ericsson and Simons pointed out that this incompleteness does not reduce the value of the data

collected through TAPs. For one thing, the reported data should be sufficient to infer the nature of the unreported processes. For another, without such an approach that provides direct evidence about cognitive operations, the mental activities might stay invisible for research purposes.

To address reactivity, Ericsson and Simon (1993) specified that if participants are only asked to verbally report cognitive processes stored in short-term memory (either coded in verbal form or not) without being required to explain and justify their mental processes, TAPs do not alter the performance, thus leaving the processes and products unmodified. However, it is pointed out that transforming processes that are not verbally coded in the first place (such as visual information) to verbal codes might slow down the performance to various degrees.

CHAPTER 3

METHODS AND MATERIALS

In order to investigate the research questions listed in Chapter 1, quantitative textual analysis and qualitative process analysis were employed. The quantitative textual analysis examines whether and how linguistic features of TOEFL iBT essays vary with task type, essay scores, and academic experience of test takers. The qualitative process analysis aims to find out whether and how writing processes vary with task type, essay scores, and academic experience of test takers. In the following sections, the research design for the quantitative textual analysis is presented first followed by that of the qualitative process analysis.

Quantitative Textual Analysis

The following section provides detailed information of the methods and materials used for the quantitative textual analysis component of the study. The information is presented in the order of the data, the instrument, and the statistic analysis.

Data

The data for textual analysis was provided by Educational Testing Service (ETS). The data includes two sets of computerized integrated and independent essays. The first set comes from a TOEFL iBT administration in 2006. A sample of 240 test takers' essays was collected across the two tasks: independent and integrated. The second set of data came from an administration in 2007. It also includes integrated and independent essays produced by 240 test takers. All the essays were graded by ETS-trained raters (see Appendix A for the scoring rubrics). A final score is available for each of the integrated and the independent essays. In addition to the computerized essays, the task prompts (including the source texts for the integrated writing task)

and the background information of the test takers were also provided by ETS. Details of the data will be presented in Chapter 4.

Instrument: Coh-Metrix

In this study, Coh-Metrix was used to generate scores for the linguistic features of the TOEFL iBT essays in terms of their basic text information, lexical sophistication, syntactic complexity, and cohesion. Why these features were selected and how they were measured and reported by Coh-Metrix will be presented in Chapter 4, where the analysis is described in detail.

Data Analysis

To answer the research questions related to linguistic features of the integrated and independent essays, a series of statistic analyses on the scores generated through Coh-Metrix analysis were performed. Discriminant (Functional) Analysis (DA) was used to address research question 1 regarding linguistic differences between the two types of essays. Regression analysis was used to answer research question 2 about linguistic differences across score levels. One-way ANOVA was used to answer research question 3 about whether linguistic features vary along with academic experience of the test takers. Only brief information about the statistic analyses is presented in the following section. The analyses will be discussed in greater detail in Chapter 4 and Chapter 5 respectively.

Research Question 1

Research question 1 focuses on whether linguistic features vary with task type (i.e., the integrated and the independent writing tasks). A MANOVA and a DA were conducted on the essays at all proficiency levels (defined by the scores assigned) to identify possible linguistic differences between the integrated and the independent essays. However, as mentioned previously, Coh-Metrix does not report on language errors, and essays with lower scores might

not meet the task requirement and contain many sentence level mistakes that might mislead the computational analysis and the following statistical analysis. With this concern in mind, a MANOVA and a DA was also performed on the essays with scores no lower than 3.5 points (out of 5 points) to further clarify the linguistic differences between the integrated and the independent essays. A preliminary analysis, which will be further described in Chapter 4, indicated the 2006 and the 2007 data sets were similar to each other in terms of the differences identified between the two types of essays. Since the data that was collected in 2007 contained more highly rated essays than the 2006 set, the following study focuses on reporting the results of the 2007 data set. The 2006 data set was kept for supplementary analysis.

Because DA is a statistical tool that has only been introduced to the analysis of L2 writing recently, the following paragraphs will focus on describing DA as well as the rationale for choosing this particular tool.

DA is a supervised classification algorithm. The term "supervised" refers to the fact that in DA, the classes (i.e., groups of cases) are predefined or already in existence (Jarvis, in press). DA is often used to test whether there are recognizable patterns associated with the predefined classes and whether these patterns are powerful enough to predict group membership of future cases. When used in textual analysis, DA can uncover whether linguistic features fed into the program are significant indicators of the text classes (different groups of texts). If particular patterns can be identified with each class of texts, a significant model can be constructed via DA. The model can then be applied to texts whose membership is withheld and predict which class they belong to in order to determine the predicting accuracy of the established model.

Therefore, DA can not only help to illustrate whether there are linguistic differences between different text classes but also to verify whether these differences are powerful enough to

predict the group membership of texts when their group information is not revealed to the model. In fact, DA has been used with success in studies that have sought to distinguish different text types including English essays produced by writers of different native languages (e.g., Crossley & McNamara, in press b) and essays by writers at different proficiency levels (Crossley & McNamara, in press a) .

This study conducted a series of stepwise DA to examine whether there are linguistic differences between the integrated and the independent essays. In this case, the redefined classes are the two types of essays (integrated and independent essays). Furthermore, DA was used to test whether the series of Coh-Metrix indices identified the group membership of the essays from the data set with accuracy.

Research Question 2

The second research question concerns how linguistic features relate to essay scores within each task type. To address this question, regression analyses were used to investigate the predictive ability of linguistic features to explain the variance in the scores of the integrated and the independent essays. Selected Coh-Metrix indices were regressed against the holistic scores of the 480 essays collected in 2007. The criterion used to choose which Coh-Metrix indices will be described in details in Chapter 4.

Research Question 3

Research question 3 is about the relationship between linguistic features and academic experience of test takers. In answering this research question, only a subset of the 2007 essays was compared through one-way ANOVA. More specifically, the test takers who applied to graduate programs were compared with those who applied to undergraduate programs in terms of the linguistic features used in their essays. Forty eight of the 240 test takers indicated that they

took the test to become a graduate student while 51 reported that they took the test to enroll in undergraduate programs. The rest of the test takers either took the test to enroll in summer programs or did not specify their reasons to take the test. To further confirm the finding, an independent *t*-test was also conducted on the essay scores of the 48 and 51 test takers to see whether the same picture emerged in terms of the scores.

Qualitative Process Analysis

Qualitative process analysis focuses on writing processes elicited by the integrated and the independent writing tasks. The research questions are about whether writing processes vary with task type, essay scores, and academic experience of test takers. Think-aloud protocols (TAPs) were used to elicit data regarding writing processes. A pilot test of three ESL participants from Georgia State University (GSU) was conducted to determine the feasibility of the research design. The three participants were purposively chosen with varying English proficiency: the first was a pre-matriculated ESL writer from the Intensive English program, the second was a matriculated undergraduate student, and the third one was a matriculated graduate student. Based on the observations and interviews of the three participants, changes were made to ensure that the data collection procedure is efficient and effective. The following sections begin with introducing the participants, the instruments, and the data collection procedures.

Participants

A total of 20 ESL students participated in the writing process component of the study. The participants were enrolled at GSU in the spring semester of 2011. They were recruited through posted flyers (see Appendix B) on campus as well as announcements of several ESL writing instructors in their classes. In the flyers, it was clearly stated that participation was completely voluntary and would have no impact on any class evaluations of the participants. The

flyers also provided a detailed description of the study including the research purpose and the tasks the participants were expected to perform. It was also stated in the flyers that $50 would be rewarded for participation in the study.

The participants were selected based on several criteria. First of all, they all had to be matriculated students at GSU. Students enrolled in the Intensive English Program were not considered for the study due to the following reasons. The first reason is that, according to the pilot study, participants with limited language proficiency, especially listening ability, tended to solely reply on the reading passage while ignoring the listening material. Only matriculated students were selected also because limited language proficiency might hinder the think-aloud writing processes because lack of speaking ability not only adds to the cognitive load but also prevents the participants from successfully completing the think-aloud writing tasks in English. The second criterion is that to answer research question 6 (whether writing processes vary with academic experience of test takers), the participants should be evenly divided between graduate and undergraduate students. I decided not to accept more than three participants from any department or linguistic background at both graduate and undergraduate levels. The disciplinary and linguistic backgrounds were controlled for two reasons: a) the more diverse the disciplinary and linguistic backgrounds the participants are from, the more representative they are, and b) according to previous research, either disciplinary background or linguistic background exerts influence on writing processes adopted by writers (Cohen, 1998; Friend, 2001). As for the third criterion, all the participants had to be in their first year studying in the United States and had not received their previous degree(s) in a country with an English medium of instruction. This was specified on the account that with limited educational experience in the U.S., the participants can better represent the target population of TOEFL iBT test takers.

The plan for participant recruitment was approved by the Human Subjects Committee Institutional Review Board (IRB) at GSU (see Appendix C). IRB also approved the research methods and documents described in the following sections of this chapter. Detailed information about the participants will be presented in the results section of qualitative process analysis in Chapter 4.

Instruments

The data for the qualitative component of the study was collected using TAPs. A set of documents were employed in the data collection procedures, which include a) a background questionnaire, b) a TAP training sheet, c) post-task questionnaires on the integrated and the independent writing sessions, and d) a semi-structured interview. I will first describe the procedures taken to ensure the TAP data quality and then provide more information on each of the documents used.

Because of the potential distraction and unnaturalness associated with verbally reporting cognitive operations in TAP (as discussed in Chapter 2), the following strategies were adopted in collecting the TAP data. First of all, in think-aloud writing sessions, any possible interaction between the researcher and the participants was avoided or reduced to a minimum level. This was done because social interaction is likely to invite modification on the thinking processes or the report of thinking processes of the participants (Ericsson & Simons, 1993; Swarts et al., 1984). Another major concern associated with TAPs is that individuals' tendency to be silent while engaging in composing, especially with L2 writers conducting thinking aloud in English (Sasaki, 2000). To address this issue, training sessions were provided to each participant to illustrate what was expected from them. The training was conducted all in English. However, in order to avoid overlearning, different tasks were used in the training session (Ericsson & Simons,

1993). Moreover, when participants stop verbalizing their mental processes, reminders were given to urge them to keep reporting their mental processes (Ericsson & Simons, 1993; Plakans, 2007; Stratman & Hamp-Lyons, 1994).

The subsequent sections introduce each of the documents that was used in collecting the TAP data.

Informed Consent Form

At the very beginning of the think-aloud writing sessions, each participant was given a copy of the informed consent form (Appendix D) to sign. In the form, information about the purpose, procedure, potential risks and benefits of the study is presented. In addition, the form also contains information on voluntary participation and withdrawal and confidentiality concerning the participants' involvement in the study.

Background Questionnaire

The background questionnaire (Appendix E) elicited demographic information from the participants. The information includes their gender, home country, native language, academic status, English writing courses and writing experience, and previous TOEFL scores if they had taken the test.

TAP Training Sheet

For the training session, each participant received an instruction sheet, which includes a written explanation of the aim of the study and details of what they were expected to do. The training sheet is presented in Appendix F. The first section of the training sheet (the research purpose and the expectations) was read aloud to the participants. Then I demonstrated TAPs with a picture comparison task. After that, the participants were required to perform think-aloud on a different picture comparison task. At the end of the training session, the participants used TAPs

to write an email to make sure that they had a clear understanding of what they were expected to perform during writing. It was also made clear to the participants that I would remind them if no verbal report of their mental processes was made for a period longer than 20 seconds.

Post-task Questionnaires

Two post-task questionnaires were also presented to the participants: one for each writing task. These questionnaires aim to elicit information about the participants' perception of and experience with the two think-aloud writing sessions. The questionnaires for the integrated and the independent writing tasks can be found in Appendix G and Appendix H respectively.

Semi-structured Interview

A semi-structured interview was also included. In the interviews, when necessary, video-tapes were replayed to refresh the participants' memory of their writing sessions. The interview questions target at the reasons for their interpretation of and experiences with the think-aloud writing sessions and their particular writing behaviors. In addition, if vagueness or unclearness occurred in the verbal reports, the participants were invited to view their tapes immediately after the writing session to clarify the ambiguity. The sample questions contained in the semi-structured interview are outlined in Appendix I.

Data Collection

The participants in this study performed the think-aloud writing sessions on a one on one basis. The writing tasks were the same as the ones used in the quantitative section of the proposed study (details of the tasks are presented in Chapter 4). After signing the consent form, each participant filled out the background questionnaire about their demographic information. Prior to the real think-aloud writing sessions, training on TAPs was provided using the TAP training sheet. After the participants indicated that they had no questions with TAPs, they were

given the real writing tasks. Each participant performed two think-aloud sessions in the order of the integrated writing task and then the independent writing task. The order was the same as that in TOEFL iBT. Since verbalization adds to the cognitive demands of the task, the time constraint was not emphasized in carrying out the think-aloud writing sessions (Plakans, 2007). However, it was made clear to the participants that they should aim at their best possible performance so as to approximate their writing acts in a testing context as much as possible. During the think-aloud writing sessions, the participants' performance was video recorded. I took field notes while observing the participants. However, my involvement in the writing sessions was limited to giving reminders when the participants stopped verbalizing their cognitive processes. After completing the two think-aloud writing sessions, the participants were then asked to fill out the questionnaires on their understanding and experience of the tasks. Finally, I also conducted a semi-structured interview with each of the participants for additional information.

Data Analysis

Verbal reports of the participants were transcribed verbatim. Transcriptions of the verbal reports were cross-referenced with the participants' essays to provide a clear presentation of the writing process. The transcriptions were closely examined to find patterns that evolve into writing behaviors (segments representing an idea or an action) using guidelines from sources including Ericsson and Simon (1993) and Green (1998). The coding scheme was based on previously established coding systems including the cognitive processes of writing (Hayes & Flower, 1980), checklist of writing strategies (Grabe & Kaplan, 1996), and writing processes in discourse synthesis (Spivey, 1984) and in summary writing (Yang & Shi, 2003). Based on the results yielded in the pilot study, additional categories were added to the coding schemes to

account for the particular nature of the integrated writing task—two source texts. The coding scheme used for analyzing integrated writing and independent writing is presented in Chapter 5.

I coded the writing episodes after multiple readings of the transcriptions. As a reliability check, another experienced rater, using the same coding schemes, independently coded a portion of the think-aloud data. Of the TAP data for four participants, the agreement reached was 94.7%. Discrepancy was solved through discussion among the raters, and member checking was also performed with the participants for final coding when necessary.

To answer research question 5 (whether writing processes vary with essay scores within each task), the 40 essays produced by the 20 participants were rated by two experienced ESL raters. The same rubrics provided by ETS were used. The final scores were the average of the scores given by the two raters. In case of discrepancy, the two raters reviewed the essays together before they decided on the final scores. Details of the scores will be shown in Chapter 5.

CHAPTER 4

QUANTITATIVE TEXTUAL ANALYSIS

This chapter focuses on the first set of research questions with regards to whether linguistic features of TOEFL iBT essays vary with task type, essay scores, and academic experience of test takers. As mentioned previously, the quantitative textual analysis focused on the corpus of 480 essays from the administration in 2007 while the other data set (collected in 2006) was kept for supplementary analysis. This chapter begins by describing the corpus of 2007 essays in detail. Then the statistical analyses, results, and discussion are presented for each of the three research questions about task type, essay scores, and academic experience of test takers respectively.

Data

In addition to the computerized essays, the data set provided by ETS also contains information about the task prompts, the test takers, and the essay scores. The following sections present detailed description of the tasks, the test takers, and the essays in the 2007 corpus.

Writing Tasks

In the TOEFL iBT, both the integrated and the independent tasks were performed on computers.

Integrated Writing Task

For this particular data set, the integrated writing task contained two source texts on fish farming. The reading passage focused on presenting the negative effects of fish farming while the listening material argued against each of the points listed in the reading passage. For the integrated writing task, the test takers were required to summarize how the listening passage challenges the reading passage. The writing instruction, the reading passage, and the

transcription of the listening passage are shown in Appendix J. As mentioned earlier, the scoring rubric is presented in Appendix A.

Independent Writing Task

For the independent writing task, the test takers were asked to write an argumentative essay on the importance of cooperation in today's world as compared to that in the past. The test takers were expected to use specific reasons and examples to argue for the stance that they chose. The prompt for the specific independent task used in the study is shown in Appendix K. The scoring rubric is included in Appendix A.

Test Takers

Two hundred and forty test takers responded to the integrated and the independent writing tasks described above. The test takers included both ESL and EFL learners. The age of the test takers ranged between 14 and 50 (M = 24).They were from a variety of home countries and from diverse linguistic backgrounds. Table 4.1 summarizes the number of participants sorted by their native languages.

Table 4.1 *Participants by Native Languages*

Native language	Number	Percentage
Chinese	43	17.9%
Spanish	29	12.1%
Korean	21	8.8%
Japanese	18	7.5%
Arabic	14	5.8%
German	13	5.4%
French	10	4.2%
Other languages[1]	92	38.3%
Total	240	100%

1. Other languages include all the languages with fewer than 10 test takers.

Out of the 240 test takers, 40 of them were not identifiable by gender, and among the rest, 95 were female, and 105 were male. Forty eight of the 240 test takers indicated that they took the test to be enrolled in undergraduate programs. Fifty one wanted to apply for graduate programs,

and the rest (141 test takers) took the test for other reasons or did not specify the reasons why they took the test.

Essays

Length

Comparing the 240 integrated essays with the 240 independent essays, it was found that the two types of essays are different in term of length as indicated in the task requirements. For the independent essays, the task required a minimum of 300 words while for the integrated writing task, no fewer than 225 words were expected. Table 4.2 presents descriptive information related to the text length for the two types of essays.

Table 4.2 *Descriptive Statistics of the Length of the Integrated and the Independent Essays*

Essay type	Mean	S.D.	Minimum	Maximum	Median
Integrated	312.37	77.457	85	592	315
Independent	197.12	50.834	54	388	192

Scores

Each essay was holistically scored by ETS-trained raters on a scale of 5 points using the appropriate scoring rubric. The scores of the independent essays were on average higher than those of the integrated essays. Meanwhile, Pearson correlation test shows that the two sets of scores are highly correlated at $r = .744$ ($p<.001$). Detailed information about the number of test takers at each score level together with the descriptive statistics of the scores is presented in Table 4.3. As can be seen, the scores on the integrated task were more evenly spread out than those on the independent task.

Table 4.3 *Number of Test Takers at Each Score Level and Descriptive Statistics of the Scores*

Score	Integrated	Independent
5	35	25
4-4.5	57	66
3-3.5	56	100
2-2.5	50	45
1-1.5	42	4
M	3.148	3.471
S.D.	1.308	0.910

The following sections present results and discussions for research questions 1, 2, and 3.

Research Question 1

The first research question concerns whether linguistic features vary across the two TOEFL iBT writing tasks (text-based integrated and independent writing tasks) in the corpus of 480 essays. A DA was performed to provide empirical evidence to answer this research question. Information about the variable selection for the DA model will be described first followed by the results and discussions.

Variables Selected apriori

In order to use DA to determine whether there were linguistic differences across the two types of essays, a set of Coh-Metrix indices were first fed into the program. To decide what linguistic features would be of interest to answer research question 1, I consulted earlier analyses of formal academic prose in English(e.g., Biber, 1988, 1995) and previous research that compared independent writing with integrated writing (e.g., Cumming et al., 2005, 2006; Gebril & Plakans, 2009). Based on these two lines of research, several Coh-Metrix indices, as described below, were selected apriori from the following categories: lexical sophistication, syntactic complexity, and textual cohesion to address the first and the second research question. Because lexical categories (Biber, 1988) and text length (Cumming et al., 2005, 2006) also play a role in differentiating different types of writing, basic text information indices from Coh-Metrix were

also included in the initial variable selection. Note that in each measure, there are sometimes several indices as described below. Table 4.4 lists the Coh-Metrix indices selected for the DA.

Table 4.4 *Summary of Coh-Metrix Indices Pre-selected for the DA*

Categories	Coh-Metrix measures	Number of indices	Direction*
Basic text information			
	Text length	4	/
	POS tags (lexical categories and phrases)	12	/
Lexical sophistication			
	Word length	1	+
	Word hypernymy value	3	+
	Word polysemy value	1	-
	Lexical diversity	4	+
	Word frequency	6	-
	Word information (word concreteness, familiarity, imagability, & meaningfulness)	8	-
	Nominalizations	1	+
Syntactic complexity			
	Number of words before the main verb	1	+
	Number of higher-level constituents per word	1	+
		1	+
	Number of modifiers per noun phrase	1	+
	Number of embedded clauses	3	-
	Syntactic similarity		
Cohesion			
	Causality	4	+
	Connectives	3	+
	Logical operators	1	+
	Lexical overlap	8	+
	Semantic similarity (LSA and LSA/given and new)	3	+
	Tense and/or aspect repetition	4	+

* Direction refers to how Coh-Metrix index scores relate to the linguistic property they represent in theory. For instance, for the index of word length, the symbol "+" means that a high score of this index suggests a higher level of lexical sophistication.

In the following sections, the indices will be introduced with more detail in the order of basic text information, lexical sophistication, syntactic complexity, and cohesion.

Basic Text Information Indices

Coh-Metrix reports basic textual information by reporting the number of words, sentences, and paragraphs per text and the number of sentences per paragraph. In addition, Coh-Metrix also generates frequency data of 13 POS tags including different types of lexical categories and phrases. The scores of these POS tags are normalized on 1,000 words.

Lexical Sophistication Indices

Coh-Metrix evaluates lexical sophistication of a given text by calculating syllables per word, lexical hypernymy and polysemy values, lexical diversity, word frequency, word information, and nominalizations.

Syllables per Word

One way Coh-Metrix measures lexical sophistication is by counting syllables per word. The more syllables that words have in a given text, the higher the word length score is, suggesting a higher degree of lexical sophistication. Previous studies (e.g., Grant & Ginther, 2000) illustrated that writers with higher proficiency tend to use longer words with more syllables.

Hypernymy

Coh-Metrix reports hypernymy values of a given text for the words that have entries in WordNet (Fellbaum, 1998), an electronic lexical database, which provides word sense information of nouns, verbs, adjectives, and adverbs. The hypernymy values are calculated by counting the number of levels that is above a word in a conceptual taxonomic hierarchy (Graesser et al., 2004; Crossely & McNamara, in press a). For instance, the word *mower* has more hypernymy levels than *machine*. Words with more hypernymy levels tend to be more precise in signaling the intended meaning and less ambiguous than those with fewer levels

(Graesser et al., 2004). Due to this reason, a higher hypernymy value indicates a higher degree of sophistication in terms of vocabulary choice.

Polysemy

Polysemy refers to the number of senses that a word has. Therefore, polysemy scores indicate lexical ambiguity of a given text. Coh-Metrix also uses WordNet to report polysemy values. Words with high polysemy values are generally more ambiguous and thus may take longer to comprehend, especially for less experienced readers (Gernsbacher & Faust, 1991; McNamara & McDaniel, 2004). They also tend to be more frequent words (Zipf, 1945). Due to this reason, texts with high polysemy values are often lexically less sophisticated.

Lexical Diversity

Coh-Metrix estimates lexical diversity using MTLD (McCarthy & Jarvis, 2010) and D (Malvern & Richards, 1997; Jarvis, 2002) values. As mentioned previously, MTLD is calculated as the mean length of sequential word strings in a text that sustain a criterion level of lexical variation (McCarthy & Jarvis, 2010). D measures lexical diversity through a computational procedure that utilizes ideal TTR curves (McNamara & Graesser, in press). Different from traditional measures of lexical diversity such as TTR, these new measures are more reliable because they avoid the problematic correlation with text length (Crossley & McNamara, in press a). A high lexical diversity score means that the given text contains a wide range of words, thus showing more lexical sophistication.

Word Frequency

Word frequency indices show how often particular words occur in the English language. Coh-Metrix word frequency counts are primarily based on CELEX (Baayen, Piepenbrock, & van Rijn, 1993), the database from the Dutch Center for Lexical Information. CELEX consists of

word frequencies taken from the early version of the COBUILD corpus of 17.9 million words. Frequent words are normally retrieved and processed quickly in meaning construction (Rayner & Pollatsek, 1994). In L2 writing assessment research, more advanced L2 writers have been found to produce texts with less frequent words (Grant & Ginther, 2000; Reid, 1990). A high word frequency score means that the input text contains more frequent words, thus indicating less lexical sophistication.

Word Information

Word information measures report values for word concreteness, familiarity, imagability, and meaningfulness. Coh-Metrix reports these indices using human ratings of linguistic properties of words provided by the Medical Research Council (MRC) Psycholinguistic Database (Wilson, 1988). A word that refers to a tangible entity tends to have a higher concreteness score than an abstract word (Toglia & Battig, 1978); therefore, a high concreteness score indicates a low level of lexical sophistication. Familiarity signals how readily recognizable a given word is. It is important, however, to be aware that familiar words do not have to be frequent words (Crossley& McNamara, in press a). A high familiarity score shows that the words used tend to be familiar words, indicating a low level of lexical sophistication. Imagability indicates whether a word can easily evoke a mental image. A high word imagability score means that the given text includes many words that can easily be associated with mental images, thus having a low level of lexical sophistication. A word with high meaningfulness score is a word that can be associated with many other words. Therefore, a high meaningfulness score often indicates that the input text contains many words that have strong association with others words, suggesting a low level of word sophistication. All of these indices are important indicators of word knowledge of a writer (Crossley & McNamara, in press a). For each of the four indices,

Coh-Metrix reports the scores for content words only and for all words per given texts respectively.

Nominalizations

Nominalizations refer to abstract generic nouns that are derived from another part of speech via the addition of derivational morphemes (e.g., *-ment, -tion, -lity, -ness*; Biber, 1988). Similar to other indices, Coh-Metrix reports this index on a normalized scale. The higher the normalization score is, the more sophisticated the words of the given text are presumed to be.

Syntactic Complexity Indices

Coh-Metrix measures syntactic complexity using five indices including the number of words before the main verb, number of higher-level constituents per word, number of modifiers per noun phrase, syntactic similarity, and number of embedded clauses. Syntactically complex sentences are often structurally elaborated and ambiguous and have many levels of embedded constituents (Graesser, et al, 2004; Perfetti, Landi, & Oakhill, 2005).

Mean Number of Words before the Main Verb

Coh-Metrix reports the average number of words before the main verbs of main clauses in sentences. The more words there are before the main verb, the more complex the sentence tends to be structurally. Therefore, a high score of this index suggests a high level of syntactic complexity of a given text.

Higher-level Constituents per Word

By using a syntactic parser to assign tree structures to sentences, Coh-Metrix calculates the number of higher-level constituents per word in a given text. Higher-level constituents refer to sentences and embedded constituents at different phrase and clause levels. Sentences with

difficult syntactic composition tend to have a higher ratio of high level constituents per word than sentences with less complicated structure (Grasser et al., 2004).

Number of Modifiers per Noun Phrase

Within each noun phrase, Coh-Metrix counts the number of modifiers. Although modifiers are optional elements in noun phrases, they often indicate how compressed the sentence structure is and signal the density of the information (Biber & Gray, 2010). A high score of this index thus suggests that the given text is syntactically more complicated and condensed.

Syntactic Similarity

The syntactic similarity index compares the syntactic tree structures of sentences. A higher syntax similarity score means a higher degree of similarity in syntactic structure of two adjacent sentences or among all sentences within a paragraph or a text and less syntactic variation (Crossley & McNamara, in press a). Therefore, a high syntactic similarity score indicates a low degree of syntactic complexity.

Number of Embedded Clauses

Coh-Metrix also reports on the number of embedded clauses of a given text as another measure of syntactic complexity. Unlike the higher-level constituent per word index, this index only focuses on embedding at the clause level (rather than embedding at both the clause and the phrase levels) The higher the number of embedded clauses is, the more complex the syntactic structure of the given text is as compared to one mainly containing simple sentences without embedding (Graesser et al., 2005).

Cohesion Indices

Textual cohesion consists of linguistic devices that play a role in building links between ideas in a given text. It is, therefore, vital in successful processing and comprehension of texts (Grasesser, McNamara, & Louwerse, 2003; Halliday & Hasan, 1976). Coh-Metrix reports cohesion by examining causality, connectives, logical operators, lexical overlap, semantic similarity, and tense and/or aspect repetition.

Causality

Causality (causal cohesion), evidenced by causal verbs, causal particles (such as *as a result*, *because*, etc), and causal connectives, reflects the extent to which sentences are linked in a text. These linguistic devices help to create connections between sentences and ideas (Pearson, 1974-1975). Presumably, a high causality score means that the given text is cohesive with causal relationship built among the ideas.

Connectives

Connectives are mainly used to create links between ideas and clauses (Halliday & Hasan, 1976) and thus are important indicators of text organization (van de Kopple, 1985) and text cohesion. Connective indices reported by Coh-Metrix include different types of cohesion such as causal connectives (e.g., *because, so, consequently*) and logical connectives (e.g., *or, actually, if*).

Logical Operators

Logical operators (*or*, *and*, *not*, *if*, and their variants) are often frequently used in texts that express logical reasoning (Crossley & McNamara, in press a). A high frequency of logical operators suggests a high level of textual cohesion.

Lexical Overlap

Coh-Metrix reports four forms of lexical overlap between sentences: noun overlap, argument overlap, stem overlap, and content word overlap. Argument overlap focuses on nouns and reports how often nouns with common stems (including pronouns) are shared between two adjacent sentences. Stem overlap indices also focus on nouns, but they look at how often a noun in one sentence shares a common stem with other word types in another sentence without counting pronouns (Crossley & McNamara, in press a). If a text has a high lexical overlap score, the text often displays a high level of cohesion as evidenced in its word choice.

Semantic Similarity (LSA and given/new information)

Coh-Metrix utilizes LSA to report semantic similarity among text constituents (e.g. word, clauses, sentences, etc). LSA is a mathematical and statistical technique for representing deeper world knowledge based on large corpora of texts. In addition, Coh-Metrix also estimates the proportion of new information each sentence provides by using LSA. The given information, since it can be retrieved from preceding text, is less taxing on the cognitive load (Chafe, 1975) and contributes to textual cohesion. Therefore, semantic similarity is an important indicator of text cohesion and increases along with the increase in text cohesion.

Tense and/or Aspect Repetition

These indices refer to temporal cues provided in an input text. They help to construct a more coherent model of the text (Crossley & McNamara, in press a). Coh-Metrix uses tense repetition, aspect repetition, and the combination of aspect and tense repetition to measure temporal cohesion embedded in the input text.

To provide a better understanding of the Coh-Metrix analysis, in Appendix L, one sample integrated essay and one sample independent essay are presented together with their scores for each of the Coh-Metrix indices above mentioned.

Variable Selection for DA

Within-subject ANOVAs were first conducted to determine which of the pre-selected Coh-Metrix indices show significant differences between the two task types. The independent variable was the task type, and the Coh-Metrix indices related to basic text information, lexical sophistication, cohesion, and syntactic complexity were used as dependent variables. Since the Coh-Metrix indices pre-selected were informed by related theories and empirical studies, type one error (an error in which it is falsely believed that a difference exists) was controlled in the ANOVAs. The results of the ANOVAs ordered by the effect size are presented in Appendix M for future reference.

As previously mentioned, in many cases, Coh-Metrix uses different indices to estimate the same linguistic property. For instance, word frequency is reported for all words contained in a given text or for all content words in the text. If any selected measures showed significant differences between the two task types, the indices that displayed the highest effect size from that measure was selected. To ensure that the selected indices were not redundant, correlation tests were conducted to ensure that none of them correlated at $r \geq .70$ (Brace, Kemp, & Sneglar, 2006). When the r value was higher than .70, the index with lower effect size (reported by the ANOVA) was removed and replaced by the index with the next highest effect size from the same measure. The same procedure was repeated until none of the selected indices were highly correlated. Tolerance and variance inflation values (VIF) were checked for the selected indices. Tolerance value of a variable indicates the portion of variance of the variable that is not related to other

independent variables in a model. VIF is the reciprocal of tolerance (O'Brien, 2007). Checking for VIF (< 10) and tolerance values (< .1) can ensure that the selected indices did not suffer from severe multi-collinearity and were not redundant (Neter, Wasserman, & Kutner,1989; Hair, Anderson, Tatham, & Black, 1995).

The indices which survived the ANOVAs, correlation, and VIF and tolerance check were then submitted to a DA to verify whether any of them (or combination of them) are predictive of the two task types (independent and integrated writing tasks). The study then conducted a DA using 10-fold cross-validation techniques with embedded feature selection (hereafter referred to as the 10 CV set). 10-CV provides optimal reliability and efficiency in testing classification models (Lecocke& Hess, 2006). The whole set of essays was randomly divided into ten folds of 48 essays. In each fold of the analysis, one fold was withheld as a test set while the other nine folds, the training set, were used to construct a model. The model obtained from the training set was then used to classify the essays in the withheld set. This procedure was performed ten times so that each single essay was classified independently of the training set.

Results for Research Question 1

As mentioned in Chapter 3, low rated essays might contain severe sentence level mistakes that can mislead the computational and statistic analysis of the linguistic features. Due to this reason, DA was performed on the whole data set of 480 essays as well as on the subset of the essays with scores no lower than 3.5 points. The results of the whole data set will be presented first followed by those of the subset.

Results for the Whole Data Set of 480 Essays

All together, 26 indices met all the criteria and were uploaded to the DA to generate a model that could classify the essays into the two task types. Descriptive statistics of the 26 indices ordered by effect size (eta squared) are shown in Table 4.5.

Table 4.5 *Means (standard deviations), F values, and Effect Sizes for the Essays in the Total Set*

Coh-Metrix indices	Categories	Independent	Integrated	F(1,478) [a]	η2
Word concreteness (content words)	Lexical sophistication	347.302 (15.031)	414.875 (18.837)	2362.294	0.908
Number of words per text	Basic text information	312.367 (77.457)	197.125 (50.834)	860.109	0.783
Stem overlap	Cohesion	0.402 (0.187)	0.764 (0.187)	553.875	0.699
Nominalizations	Lexical sophistication	11.358 (6.166)	3.600 (2.398)	455.889	0.656
Verbs in base form	Basic text information	52.902 (16.769)	28.250 (15.882)	422.745	0.639
Word frequency (all words)	Lexical sophistication	3.227 (0.093)	3.093 (0.114)	397.899	0.625
Number of higher-level constituents per word	Syntactic complexity	0.766 (0.0368)	0.711 (0.035)	388.725	0.619
Lexical diversity	Lexical sophistication	77.411 (17.422)	52.535 (14.412)	378.055	0.613
Verbs in 3rd person singular present form	Basic text information	26.362 (11.328)	48.873 (19.045)	308.548	0.564
Personal pronoun possessive cases	Basic text information	15.378 (10.184)	3.799 (5.437)	245.773	0.507
Past participle verbs	Basic text information	13.174 (9.668)	26.966 (15.634)	211.484	0.469
Verbs in non-3rd person singular present	Basic text information	36.902 (15.936)	22.484 (13.002)	175.379	0.423
Hypernymy values of nouns	Lexical sophistication	5.464 (0.558)	5.978 (0.542)	156.551	0.396
Logical operators	cohesion	45.126 (15.841)	33.824 (13.594)	94.132	0.283
Number of paragraphs per text	Basic text information	4.83 (1.835)	3.83 (1.607)	90.111	0.274
Word meaningfulness (content words)	Lexical sophistication	423.541 (14.607)	433.820 (14.037)	72.175	0.232
Verbs in past tense	Basic text information	14.963 (10.954)	7.004 (9.763)	71.236	0.230

Causal verbs	Cohesion	23.406 (10.045)	16.816 (9.848)	62.928	0.208
Positive causal connectives	Cohesion	16.905 (9.466)	22.092 (11.328)	38.437	0.139
Polysemy values	Lexical sophistication	3.945 (0.420)	3.731 (0.463)	32.888	0.121
Tense aspect repetition	Cohesion	0.581 (0.170)	0.700 (0.283)	31.964	0.118
Positive logical connectives	Cohesion	34.008 (12.086)	39.934 (15.746)	28.446	0.106
LSA given/new information	Cohesion	0.296 (0.037)	0.310 (0.049)	17.393	0.068
Embedded clauses	Syntactic complexity	51.355 (16.526)	54.980 (17.225)	7.778	0.032
Syntactic similarity	Syntactic complexity	0.093 (0.028)	0.116 (0.038)	91.411	0.028
Prepositional phrases	Basic text information	112.084 (22.447)	116.447 (25.129)	5.187	0.021

* For all indices $p < .001$.
[a] Wilks' Lambda F value.

For the DA, the significant level for an index to be entered or to be removed from the model was set at .05. For the total set, the DA retained 15 out of the 26 indices as significant predictors while the other 11 were removed. For the 10 CV set, the DA retained 14 of the same indices retained in the total set as significant predictors. The other index that was retained in the total set was only retained in four of the folds but not in the other six. For the other 11 indices removed in the total set, eight of them were not retained in any of the 10 folds, two of them were retained in only one fold, and one of them was retained in two folds. The selected indices and their retention information for both the total set and the 10 CV set are shown in Table 4.6.

Table 4.6 *Index Retention in Total Set and 10 CV Set in the Whole Data Set*

Coh-Metrix indices	Retained in the total set	Number of folds retained in the 10 CV
Word concreteness (content words)	+	10
Total number of words per text	+	10
Stem overlap	+	10
Nominalizations	+	10
Number of higher-level constituents per word	+	10

Personal pronoun possessive case	+	10
Logical operators	+	10
Word meaningfulness (content words)	+	10
Verbs in past tense	+	10
Causal verbs	+	10
Positive logical connectives	+	10
Verbs in 3rd person singular present form	+	8
Tense aspect repetition	+	8
Verbs in base form	+	7
Prepositional phrases	+	4
Positive causal connectives	-	2
Past participle verbs	-	1
LSA given/new information	-	1
Word frequency (all words)	-	0
Lexical diversity	-	0
Verbs in non-3rd person singular present form	-	0
Hypernymy values of nouns	-	0
Number of paragraphs	-	0
Polysemy values	-	0
Embedded clauses	-	0
Syntactic similarity	-	0

An estimation of the accuracy of the analysis was made by plotting the correspondence between the groupings (the task types) using both the total set and the 10 CV set. The classification results from the total set and the 10 CV set are reported in Table 4.7.

Table 4.7 *Predicted text type versus actual text type results from total set and 10 CV set in the Whole Data Set*

Actual text type		
Total set	Independent	Integrated
Independent	240	0
Integrated	0	240
10 CV set	Independent	Integrated
Independent	240	0
Integrated	0	240

The classification results demonstrate that the model correctly allocated 480 of the 480 essays in the total set ($df = 1$, $n=480$, $\chi^2=480.00$, $p < .001$) with a classification accuracy of 100%

(chance for this analysis is 50%). The reported Kappa =1, indicates a perfect agreement between the actual essay classification and the predicted essay classification for the total set. The DA results of the 10 CV set also correctly allocated 480 essays of the 480 essays (df = 1, n=480, χ^2= 480.00, p < .001) for an accuracy of 100% (chances for this analysis is also 50%). The reported Kappa = 1, indicates a perfect agreement between the actual essay classification and the predicted essay classification of the 10 CV set. The 100% predicting accuracy demonstrates that the precision and recall values of the model for either the total set or the 10 CV set are 1. Recall scores are calculated by tallying number of hits over the number of hits and misses. Precision refers to the number of correct predictions divided by the sum of the number of correct predictions and false positives (Crossley & McNamara, in press b).

In summary, the DA results of the whole data set showed that 14 Coh-Metrix indices were retained in the majority of the 10 CV set and can significantly predict the essay types. As part of the ANOVAs (see Table 4.5), a comparison was conducted to show the direction of the differences between the integrated and the independent essays for each of the selected Coh-Metrix indices. The results are reported in Table 4.8 for the 14 indices. In the table, the index was listed under the essay type where its score was significantly higher than in the other type.

Table 4.8 *Predictive Indices for Task Types in the Whole Data Set (Listed in the Order of Effect Size)*

Independent essays		Integrated essays	
Coh-Metrix indices	Categories	Coh-Metrix indices	Categories
Number of words per text	Basic text information	Word concreteness (content words)	Lexical sophistication
Nominalizations	Lexical sophistication	Stem overlap	Cohesion
Verbs in base form	Basic text information	Verbs in 3rd person singular present form	Basic text information

Number of higher-level constituents per word	Syntactic complexity	Word meaningfulness (content words)	Lexical sophistication
Personal pronoun possessive case	Basic text information	Tense aspect repetition	Cohesion
Logical operators	Cohesion	Positive logical connectives	Cohesion
Verbs in past tense	Basic text information		
Causal verbs	Cohesion		

DA Results for the Subset of the Higher Rated Essays

In order to keep the within-subject design, the essays written by the same test takers who scored no lower than 3.5 points on both tasks were selected as the higher rated essays. All together, there were 106 test takers who met the requirement.

The same statistical analyses were repeated for these 212 essays. Within-subject one way ANOVAs were first conducted using the reported scores of the selected Coh-Metrix indices as the independent variables and the task type as the dependent variables. Following the same procedure in selecting variables, 23 Coh-Metrix indices were uploaded to the DA to generate a model that can classify the 212 higher rated essays into the two task types. Table 4.9 provides descriptive statistics of the 23 variables ordered by the effect size (eta squared).

Table 4.9 *Means (standard deviations), F values, and Effect Sizes for the Higher Rated Essays*

Coh-Metrix indices	Categories	Independent	Integrated	*F(1,210)*	η^2
Word concreteness (content words)	Lexical sophistication	349.801 (13.449)	419.880 (14.841)	1663.316	0.941
Nominalizations	Lexical sophistication	14.110 (6.376)	4.380 (2.584)	269.773	0.720
Lexical diversity	Lexical sophistication	82.593 (17.034)	54.481 (11.776)	268.716	0.719
Semantic similarity (sentence to sentence)	Cohesion	0.178 (0.062)	0.291 (0.066)	245.94	0.701
Verbs in base form	Basic text information	48.561 (14.511)	22.482 (13.281)	245.272	0.700

Verbs in 3rd person singular present	Basic text information	25.899 (10.971)	50.821 (18.233)	177.342	0.628
Number of modifiers per noun phrase	Syntactic complexity	0.787 (0.148)	1.003 (0.162)	165.152	0.611
Personal pronoun possessive case	Basic text information	14.559 (9.418)	3.427 (4.519)	117.849	0.529
Past participle verbs	Basic text information	17.790 (9.796)	33.592 (14.302)	108.092	0.507
Verbs in past tense	Basic text information	16.404 (10.800)	5.214 (8.342)	80.883	0.435
Tense aspect repetition	Cohesion	0.749 (0.111)	0.871 (0.102)	75.553	0.418
Verbs (non-3rd person sg. present)	Basic text information	30.331 (12.319)	18.449 (11.144)	58.570	0.358
Hypernymy values (nouns)	Lexical sophistication	5.691 (0.503)	6.154 (0.376)	57.861	0.355
Logical operators	Cohesion	46.208 (15.543)	33.294 (12.493)	54.926	0.343
Word meaningfulness (content words)	Lexical sophistication	422.682 (13.206)	434.229 (12.007)	49.756	0.322
Nouns (plural)	Basic text information	77.098 (21.721)	60.404 (18.310)	39.550	0.274
Ratio (causal particles to causal verbs)	Cohesion	0.683 (0.556)	1.236 (0.860)	39.680	0.274
Syntactic similarity	Syntactic complexity	0.094 (0.023)	0.110 (0.029)	34.710	0.248
LSA given/new	Cohesion	0.299 (0.034)	0.318 (0.039)	26.268	0.200
Gerund or present participle verbs	Basic text information	14.278 (9.697)	19.176 (10.156)	15.910	0.132
Embedded clauses	Syntactic complexity	47.062 (14.177)	53.729 (16.294)	14.740	0.123
Number of words before the main verb	Syntactic complexity	5.437 (2.029)	4.798 (1.890)	8.156	0.072
Polysemy values	Lexical sophistication	3.844 (0.338)	3.706 (0.423)	7.768	0.069

*For all indices, $p < 0.001$

For the total set, the DA retained seven out of the 23 Coh-Metrix indices as significant predictors while the other 16 were removed. For the 10 CV set, the DA retained the same variables retained in the total set as significant predictors. For the other 16 indices removed in the total set, 13 of them were not retained in any folds, one of them was retained in one fold, one of them was retained in two folds, and one of them was retained in three folds. The selected Coh-Metrix

indices and their retention information for both the total set and the 10 CV set are shown in Table 4.10.

Table 4.10 *Index Retention in Total Set and 10 CV Set for the Higher Rated Essays*

Coh-Metrix indices	Retained in total set	Number of folds retained in 10 CV
Word concreteness (content words)	+	10
Nominalizations	+	10
Verbs in base form	+	10
Personal pronoun possessive case	+	10
Verbs in past tense	+	10
Word meaningfulness (content words)	+	10
Embedded clauses	+	10
Nouns (plural)	-	3
Lexical diversity	-	2
LSA given/new	-	1
LSA (sentence to sentence)	-	0
Verbs in 3rd person singular present	-	0
Number of modifiers per noun phrase	-	0
Past participle verbs	-	0
Tense aspect repetition	-	0
Verbs in non-3rd person singular present	-	0
Hypernymy values of nouns	-	0
Logical operators	-	0
Ratio of causal particles to causal verbs	-	0
Syntactic similarity	-	0
Gerund or present participle verbs	-	0
Number of words before the main verb	-	0
Polysemy values	-	0

Table 4.11 shows the estimation of the accuracy of the DA for the 212 essays.

Table 4.11 *Predicted Text Type vs. Actual Text Type Results from Total Set and 10 CV Set in Higher Rated Essays*

Actual text type		
Total set	Independent	Integrated
Independent	106	0
Integrated	0	106

10 CV set	Independent	Integrated
Independent	106	0
Integrated	0	106

The classification results demonstrate that the model correctly allocated 212 of the 212 essays in the total set (df = 1, n=212, χ^2=212.00, p < .001) for a classification accuracy of 100% (chance for this analysis is 50%). The reported Kappa = 1, indicates a perfect agreement between the actual essay classification and the predicted essay classification for the total set. The DA results of the 10 fold set also correctly allocated 212 essays of the 212 essays (df = 1, n = 212, χ^2 = 212.00, p < .001) with an accuracy of 100% (chances for this analysis is also 50%). The reported Kappa = 1, illustrates a perfect agreement between the actual essay classification and the predicted essay classification of the 10 CV set. Again, either for total set or for the 10 CV set, the precision and recall values of the model are 1.

To sum up, for the subset of the higher rated essays, all together seven Coh-Metrix indices were retained in each fold of the 10 CV set and could significantly predict the essay types. Table 4.12, drawing on the descriptive statistics of these indices (see Table 4.9), demonstrates in which essay type the 7 indices showed a significantly higher score.

Table 4.12 *Predictive Indices for Task Types in the Higher Rated Essays (Listed in the Order of Effect Size)*

Independent essays		Integrated essays	
Coh-Metrix indices	Categories	Coh-Metrix indices	Categories
Nominalizations	Lexical sophistication	Word concreteness (content words)	Lexical sophistication
Verbs in base form	Basic text information	Word meaningfulness (content words)	Lexical sophistication
Verbs in past tense	Basic text information	Embedded clauses	Syntactic complexity
Personal pronoun possessive case	Basic text information		

Discussion for Research Question 1

Whole Data Set

When the whole data set (480 essays collected in 2007) was considered, the DA results demonstrated that the two different essay types (integrated vs. independent) can be predicted perfectly based on the linguistic differences that exist between the two groups. Certain linguistic features related to lexical sophistication, syntactic complexity, cohesion, and basic text information were shown to be able to significantly predict essay group membership of the essays under investigation based on their task types. The DA model actually performed with 100% accuracy in classifying the two types of essays. This finding not only illustrated that there are particular linguistic features associated with each of the task types but also demonstrated that these features are powerful enough to classify the essays into their specific task type. This study, therefore, yielded empirical evidence to substantiate the claim that the two writing tasks elicit different writing performance in terms of writing outcome (Huff et al., 2008). It indicates that adding the integrated writing task diversifies and improves the measurement of academic writing ability (Cumming et al., 2005, 2006), thus lending evidence to the rationale for the concurrent use of the two tasks in a writing test.

As listed in Table 4.8, the significant linguistic features for predicting the essay membership were primarily at the word level (word concreteness, word meaningfulness, nominalizations, verbs in base form, past tense, and in 3rd person singular present tense, and personal pronoun possessive cases). Other linguistic features related to cohesion (stem overlap, tense aspect repetition, positive logical connectives, logical operators, and causal verbs), text length, and syntactic complexity (mean number of higher-level constituents per word) were also included as significant predictors in the DA model. Below is a discussion of these linguistic

features in relation to the manner in which they help characterize the independent and the integrated essays.

Independent Essays

The DA illustrated that the independent essays, in comparison to the integrated essays, had significantly more frequent use of verbs in past tense and in base form. The high frequent use of verbs in past tense suggests that the test takers used their previous experience or world knowledge in the past in arguing for their stance taken. Verbs in base form refer to uninflected verbs used in imperative, infinitive, subjunctive mood (e.g., after verbs such as *suggest*, *insist*, etc) and verbs used after auxiliary and causative verbs (e.g., *make*, *help*, etc). It is important to notice that verbs in base form do not refer to uninflected verbs after 1st, 2nd, and plural subjects, which are categorized as verbs in non-3rd person singular present form. A closer examination of the independent essays revealed that many of the uninflected verbs in base form are grammatical mistakes. This is especially true in essays with lower scores. For instance, one independent essay that received a score of 1.5 contained a total of 24 verbs in base form, and seven of them are uninflected verbs that are actually grammatical errors. An example would be:

In conclusion, **communicate** is most important with this ages (Independent 20073055).

When they were not errors, verbs in base forms were often found to be employed by the test takers in arguing for their stance (e.g., to provide purposes and reasons, to list limitations that people face or choices that they have to make, etc). The following examples illustrate how the verbs in base form help to achieve this purpose.

> To **build** the advanced technology, it really requires the ability to **cooperate** well with others. (Independent 20073094)

> Most of time, one man can not **do** everything and everybody has to work together. (Independent 20073083)

In addition, the independent essays were also characterized by frequent use of personal pronouns in possessive case. Through analyzing the essays, it was found that the test takers used these possessives to provide supporting examples and/or to involve the readers. Two examples are provided below to illustrate:

> You may live next to **your** neighbor for several years, but you don't even know **his** or **her** name. (Independent20073287)

> I often hear people saying that **our** world has become much more individualistic and selfish than it used to be. (Independent20073186)

As shown in the two examples provided, this particular linguistic device demonstrates not only a high degree of involvement of the writers and but also a direct engagement with the readers. Through the use of personal pronoun possessives, the independent essays seem to display a personally involved and interactional style of writing, a characteristic typical of conversational registers (Biber, 1988, 1995; Connor & Biber, 1988).

Frequent use of logical operators and causal verbs also helped to characterize the independent essays when compared with the integrated essays. As linguistic features that demonstrate textual cohesion, these two devices helped the test takers to signal causal relationships between clauses and to express logical reasoning. This can be seen in the following two examples. The first example illustrates how logical operators were used to express reasoning. The second example shows how causal verbs were employed to indicate logical reasoning of the writer.

> **If** you cooperate with others you will get in return the respect of working together in group, **and** most importantly experience of working with a variety of people. (Independent20073275)

> Technological progress not only **increased** the effectiveness of the manufacturing but also **decreased** the dependancies of each employees, and also **reduced** the importance the ability to cooperate with each other. (Independent20073253)

Considering that the independent writing task asked for an argumentative essay, it is not difficult to understand the prevalence of logical operators and causal verbs. As Halliday (1994) pointed out, linguistic devices that signal causal and logical semantic relationships between actions and claims help to construct arguments by establishing logical reasoning and providing cause and proof for events.

The independent essays were also found to be syntactically more complex than the integrated essays in terms of the number of higher-level constituents per word. To fully understand the syntactic reality of the resultant writing in the two task conditions, the other three indices related to syntactic complexity (mean number of words before the main verb, number of modifiers per noun phrase, and embedded clauses) were also examined. As the ANOVA results (see Appendix M) demonstrated, in terms of number of higher-level constituents and number of words before the main verb, the independent essays were structurally more complex than the integrated essays. On the other hand, the integrated essays showed a significantly higher score on the indices of the number of modifiers per noun phrase and the number of embedded clauses (The higher number of embedded clauses is mainly associated with the source citing behaviors in the integrated writing). This finding suggests that the integrated essays were structurally more compressed (mainly relying on noun phrases to carry information) rather than more elaborated and syntactically complex, which is more similar to the style of academic writing in general (Biber & Gray, 2010).

The results also exhibited an unexpected pattern. As mentioned previously, the integrated essays were hypothesized to be more similar to academic writing assigned at the tertiary level of education than independent writing. Given that nominalizations are one of the characteristics of academic writing (Biber, 1988), it would be reasonable to expect them to be more prevalent in

integrated writing than in independent writing. Interestingly, the DA results did not support that, and there was actually a significantly more frequent use of nominalizations in the independent essays than in the integrated writing task. A closer examination of the independent essays showed that the use of nominalizations were primarily due to the prompt effect. As the prompt is related to "cooperate," words such as *cooperation, competition, advancement, communication, globalization* were used very regularly in the essays. To confirm whether this difference was prompt-specific, nominalizations were examined in the 2006 data set of integrated and independent essays on different topics. With that set of data, through ANOVA, it was found that the integrated essays used significantly more nominalizations than the independent essays with a moderate effect size ($F=14$, $df=1$, $\eta^2=.055$).

On a final note, the independent essays were also found to be significantly longer than the integrated essay. More time was given to the task, and it was specified in the task instruction that the independent essays should be at least 300 words long (being 75 words longer than the integrated essays as required). The resultant independent essays were, therefore, significantly longer than the integrated essays.

Integrated Essays

As illustrated in Table 4.8, the integrated essays, in comparison with the independent essays, were first of all characterized by the frequent use of concrete and meaningful words. This might be related to the fact that in the integrated writing, the writing content is highly controlled (as provided in the source texts). To meet the expectation of the task, the test takers should not add personal opinions but simply report what they extracted from the source texts. Furthermore, compared with the independent task, the topic of the integrated task tends to be more concrete and specific. This might be due to the reason that in the integrated task, it is possible to ask test

takers to write about more specific topics because the information can be found in the source texts. In contrast, in the independent writing task, test designers cannot count on every test taker to know enough about specific topics, so topics in this task tend to be more general. Due to these reasons, the integrated essays contained significantly more words that are concrete and have stronger associations with other words and concepts. Frequent use of the concrete words allows for less dependence on the context cues in meaning construction of vocabulary items. Thanks to the strong association with other words and concepts, words with high meaningfulness scores can "facilitate the comprehension of new vocabulary words and developing ideas that are not context dependent" (Crossley & McNamara, 2009, p. 130). Therefore, this study provided evidence that the integrated essays were more context independent than the independent essays in terms of the word choice. Examples from the integrated and the independent essays are provided below. In the first example from an integrated essay, the words *fish*, *meal*, and *eat*, are highly concrete and meaningful words. Successful detecting the intended meaning of these words does not require other cues provided in the context. In the second example from an independent essay, the words *way*, *know*, and *do* have low concreteness and meaningfulness scores. To know exact meaning of these words, readers seem to have to draw on more information from the context either from the preceding or following sentences.

> **Fishmeal** is made from **fish**, which human is not able to **eat**. (Integrated20073246)
>
> They can **do** things the best **way** and other people do not **know** as much as they **do**. (Independent20073275)

For the same reason (writing content being highly controlled), lexical overlap was another characteristic of the integrated essays. This linguistic feature also helps test takers enhance the cohesion within the integrated essays by providing coreference between sentences,

thus making the text more comprehensible and readable (Crossley, Salsbury, McCarthy, & McNamara, 2008; Kintsch & van Dijk, 1978). An example is given below:

> Fish farming is dangerous to human health as shown in the article. It points out that too many chemicals are used, but the lecturer tells us to be realistic about the use of chemicals. The poultry, beef, and pork we consume everyday all contains many chemicals, and what we should do is to compare the value of fish with the above-mentioned kinds of food rather than being fightened of the use of chemicals in fish farming. (Integrated20073302)

In strengthening textual cohesion, the integrated essays also contained heavy use of tense aspect repetition and positive logical connectives (such as *moreover*, *all in all*, etc). Tense aspect repetition helps construct a more coherent context in terms of temporality (Crossley& McNamara, in press). The significant difference in tense and aspect repetition might also be related to the specific prompt used in the independent writing as the task explicitly required comparison between the present and the past. The use of positive logical connectives illustrated that the test takers, in constructing the integrated essays, set up more cues to signal the textual organization and created more links between ideas and clauses (Halliday & Hasen, 1976).

The integrated essays were also found to include significantly more cases of verbs in 3rd person singular present tense and fewer cases of verbs in past tense. This finding, first of all, confirms that the integrated essays were mainly recounted in the present tense. Therefore, it suggests that the test takers recounted the content extracted from the source texts as current knowledge (Hinkel, 2002). A closer examination of the integrated essays also showed that the 3rd person singular form was mainly used in a) citing the source (e.g., *the reading passage states*, *the lecturer mentions*, etc), b) describing the subject matter (fish farming), and c) structuring sentences with dummy subject *it*. Examples to illustrate are provided as follows,

> The professor refutes the points that the passage discusses. (Integrated 20073113)
>
> First, fish farming helps local species to rebound. (Integrated 20073165)

> One negative issue of fish farming is that it poses some health risks to commercially grown fish. (Integrated 20073162)

Collectively, the results of the DA on the 480 essays showed that the independent essays, in comparison to the integrated essays, were characterized by a) being argumentative as evidenced by heavy use of logical operators and causal verbs, b) being reliant on verbs to provide examples from previous experiences and to facilitate reasoning as evidenced by verbs in past tense and base form, c) being involved and interactional as evidenced by more personal pronoun possessive cases used, and d) being structurally more complex but not compressed(Biber & Gray, 2010) with higher scores in number of words before the main verb and number of higher-level constituents per word but a lower score in number of modifiers per noun phrase.

On the other hand, the integrated essays seemed to place more emphasis on showing organizational cues in text construction and using lexical and tense and aspect repetition to build cohesion. As compared to the independent essays, the integrated essays were characterized by the more frequent use of verbs in 3rd person singular present tense and a larger number of modifiers per noun phrase, which indicated a more detached way of writing and an informational prose style (Biber, 1988). At the lexical levels, the integrated writing was also marked by heavy use of concrete words and meaningful words. This finding might be an artifact of integrated writing as the writing content was highly controlled. Additionally, it also showed that the integrated writing tends to be more context-independent as compared to the independent writing, another characteristic of formal, academic writing (Crossley & McNamara, 2009).

Before moving on to the discussion of the results of the higher rated essays, one question still remains regarding the DA results of the whole data set. The question is whether the finding that different linguistic patterns were associated with the integrated and the independent essays

was only restricted to the 2007 data set under investigation. To answer this question, a DA was performed on the other data set that was collected in 2006. Because detailed information about the analysis of the 2006 data set is beyond the scope of this study, only the DA results based on the total data set of the 2006 essays and the corresponding 10 CV sets are presented in the following tables. Table 4.13 compares the predictive indices for the integrated essays across the two data sets while Table 4.14 compares the predicative indices for the independent essays across the two data sets.

Table 4.13 *Predictive Indices for the Integrated Essays across the 2007 and 2006 Data Sets*

2007		2006	
Categories	Coh-Metrix indices	Categories	Coh-Metrix indices
Basic text information	Verbs in 3rd person singular present form	Basic text information	Past participle verbs Noun (plural)
Cohesion	Stem overlap Tense aspect repetition Positive logical connectives	Cohesion	Stem overlap
Lexical sophistication	Word concreteness (content words) Word meaningfulness (content words)	Lexical sophistication	Word concreteness (all words) Word imagability (all words) Hypernymy values (verbs) Nominalizations

Table 4.14 *Predictive Indices for the Independent Essays across the 2007 and 2006 Data Sets*

2007		2006	
Categories	Coh-Metrix indices	Categories	Coh-Metrix indices
Basic text information	Number of words per text Verbs in base form Personal pronoun possessive case Verbs in past tense	Basic text information	-- Verb phrases Personal pronouns
Cohesion	Logical operators	Cohesion	Logical connectives

	Causal verbs		
Lexical sophistication	Nominalizations	Lexical sophistication	Word familiarity (content words) Lexical diversity
Syntactic complexity	Number of higher-level constituents per word	--	--

Compared with the DA model for the 2007 data set (see Table 4.13 & Table 4.14), it can be seen that the model for the 2006 data set is slightly different in terms of the specific indices that were included. However, the overall picture that emerged from the DA still stays the same. For instance, the integrated essays in the 2006 data set, similar to those in the 2007 data set, also leaned more towards a detached and informational prose style as evidenced by the frequent use of past participle verbs (mainly associated with passive voice) and nouns. Furthermore, the integrated essays in the 2006 data set were also characterized by the use of more concrete and specific words (as evidenced by word concreteness, hypernymy values and word imagability scores). Finally, same as the 2007 data set, lexical overlap is still one of the predictive indices of the integrated essays.

On the other hand, the 2006 independent essays, similar to their counterparts from the 2007 data set, also showed an involved manner of communication (evidenced by the significantly more frequent use of personal pronouns). In addition, the frequent use of logical connectives also indicated that the linguistic devices that signal logical relationships still play an important role to argue for the stance taken in the independent essays.

To further investigate whether the differences identified were only restricted to the 2007 data set, the model based on that data set was also used to predict the task type of the 480 essays collected in 2006. The model achieved an overall accuracy of 89.2%. Interestingly, the model classified all the independent essays correctly, and all the misclassified cases were the integrated

essays. To be specific, the model correctly detected 78.3% (188 out of 240) but misclassified 21.7% (52 out of the 240) of the integrated essays. Although beyond the scope of this study, the specific reasons for this pattern certainly deserve further investigation.

With the DA results from the 2006 data set, it can be concluded that although differences exist in the particular indices that were included in the DA models, the two types of essays overall differ from each other in similar ways. Furthermore, the model derived from the 2007 data set also achieved an overall 89.2% accuracy (chances for this analysis is 50%). The results from the supplementary analysis of the 2006 data set, therefore, suggest that the linguistic differences between the two tasks found in this study were robust and were not just restricted to the two writing prompts of the 2007 data set. Instead, the patterns identified are probably more indicative of the differences between the text-based integrated writing and the independent writing in general.

Subset of Higher Rated Essays

The DA model constructed on the higher rated essays, 212 essays with scores no lower than 3.5 points, contained only seven linguistic features. Six of them were the same features from the previous analysis of the 480 essays and patterned similarly. These six features were word concreteness, word meaningfulness, nominalizations, verbs in base form, verbs in past tense, and personal pronoun possessive cases.

As can be seen, the follow up analysis with the subset of 212 essays indicated that when quality is being controlled, the integrated and independent writing mainly differed from each other at the lexical level. The higher rated independent essays were still found to contain more instances of personal pronouns than the higher rated integrated essays. Therefore, these independent essays still demonstrated a more involved and interactional writing style (Biber,

1988). On the other hand, with the topic being specific and the content being highly controlled, the integrated writing showed a more context-independent feature in lexical choices as words that were concrete and meaningful were used more frequently than in the independent essays.

The only feature that was different from the previous analysis was the Coh-Metrix index of embedded clauses. As the ANOVA shows (see Table 4.9), the integrated essays outperformed the independent essays in this particular feature. A closer examination of the essays suggested that this feature might be due to the summary nature of the task—to document the content while giving credit to the source of the information. An example is provided below to illustrate:

> The lecture explains how fish farming can have many benefits in contrast to the disadvantages expressed on the reading passage. (Integrated 20073275)

Taken together, all the DA conducted pointed to significant differences between the integrated and the independent essays under investigation. The results showed that the independent essays were characterized by an interactional communication style while the integrated essays leaned more toward a detached and informational prose style. These differences were found to be robust across the two sets of integrated and independent topics that were examined and across the essay quality perceived as well.

Research Question 2

As for research question 2, linguistic features that predict score differences were identified within each task type and then compared across the two task types to see whether the same or different sets of features predict score differences. To study whether linguistic features vary with writing quality (determined by the essay scores) within each task type, I used Coh-Metrix to analyze the 2007 corpus of 480 scored essays. Following Whitten and Frank (2005) and Crossley and McNamara (in press a), the integrated and the independent essays were divided into training and test sets respectively. With the training sets, regressions were conducted using

the scores as the dependent variable and Coh-Metrix indices as the independent variables. The results yielded in the training sets were later extended to the test sets to determine the predictive accuracy of the predictors identified in the regressions on an independent data set for each of the task types. In the following sections, I will first introduce how the Coh-Metrix indices were selected for the regression analysis on the training sets. Then I will report the statistic analysis results of the training set and how they were extended to the test set for each of the writing tasks before presenting the discussions.

Variable Selection

Informed by related research findings on linguistic features in relation to writing quality and examination of the scoring rubrics (see Appendix A), Coh-Metrix indices related to lexical sophistication, syntactic complexity, and cohesion in addition to basic text information indices were first selected to examine the role they play in determining the essay scores. The same corpus of 480 essays (collected in 2007) was used for this analysis.

Following a 67/33 split (Whitten & Frank, 2005), the 240 integrated essays were first randomly divided into a training set of 160 essays and a test set of 80 essays. An initial analysis was conducted on the training set to decide which of the preselected Coh-Metrix indices were important in explaining the essay scores. Pearson correlations were used to compare the essay scores to the reported Coh-Metrix indices. The Coh-Metrix indices were selected by the strength of r value. The variables with the highest r values within each class of measure were first selected. Another correlation test was then conducted among those selected indices to ensure that no redundant indices were included in the later regression analysis. Among each pair of indices that were highly correlated with each other ($r \geq .70$), the one with the lower correlation r value with the essay score was removed and replaced with the index from the same measure that had

the next highest correlation *r* value. The correlation test was repeated until no selected Coh-Metrix indices were highly correlated with each other.

A stepwise regression analysis was then conducted on the training set of 160 randomly selected integrated essays to examine which of the selected indices were significantly predictive of the integrated essay scores and accounted for the largest amount of variance associated with the essay scores. The selected indices were regressed against the holistic scores for the 160essays with the essay scores being the dependent variable and the Coh-Metrix indices being the predictor variables. The derived regression model was then applied to the test sets to predict the scores of the essays and the predicting accuracy of the model was calculated for the test set. The same procedure for variable selection was followed for the 240 independent essays. Regression analysis was also conducted on the 160 randomly selected independent essays (the training set) to identify which of the selected Coh-Metrix indices significantly predict the independent scores if there are any and then the model was applied to the test set to determine its predicting accuracy with independent essay samples.

Results for Research Question 2

The regression analysis results for the integrated essays will first be presented followed by those for the independent essays.

Integrated Essays

For the training set of 160 integrated essays, 19 Coh-Metrix indices entered the regression analysis. Table 4.15 presents the 19 selected indices with their *r* values and *p* values in the order of the strength of the correlation.

Table 4.15 *Selected Coh-Metrix Indices for Regression Analysis of the Integrated Essays: Training Set*

Coh-Metrix indices	Categories	*r* value	*p* value
Number of words per text	Basic text information	0.513	<0.001
Word familiarity (content words)	Lexical sophistication	-0.440	<0.001
Past participle verbs	Basic text information	0.437	<0.001
Word frequency (content words)	Lexical sophistication	-0.436	<0.001
Verbs in base form	Basic text information	-0.403	<0.001
Nominalizations	Lexical sophistication	0.357	<0.001
Hypernymy values (nouns)	Lexical sophistication	0.351	<0.001
Verbs in non-3rd person singular present form	Lexical sophistication	-0.344	<0.001
Personal pronouns	Basic text information	-0.315	<0.001
Semantic similarity(LSA sentence to sentence)	Cohesion	0.296	<0.001
Number of modifiers per noun phrase	Basic text information	0.264	<0.050
Word concreteness (content words)	Lexical sophistication	0.225	<0.050
Number of sentences per text	Basic text information	0.218	<0.050
Noun overlap	Cohesion	0.217	<0.050
Verbs in 3rd person singular present form	Basic text information	0.194	<0.050
Gerund or present participle verbs	Basic text information	0.186	<0.050
Tense repetition	Cohesion	0.174	<0.050
Prepositional phrases	Basic text information	0.168	<0.050
Verbs in past tense	Basic text information	-0.165	<0.050

The indices were also checked for outliers and multi-collinearity by examining VIF values and tolerance. All VIF values of the selected indices were found to be about 1, and all tolerance values were beyond the threshold level of .1, which indicated that the selected indices did not suffer from multi-collinearity (Menard, 1995).

With the 19 indices as the independent variables, the regression yielded a significant model, $F(1, 152) = 30.446$, $p < .050$, $r = .764$, $r^2 = .584$, adjusted $r^2 = .565$. Seven Coh-Metrix indices were included as significant predictors of the essay scores. The seven indices were: number of words per text, past participle verbs, word familiarity (content words), verbs in 3rd person singular present form, semantic similarity (LSA sentence to sentence), verbs in base form, and word frequency (content words). Descriptive statistics of the seven indices are provided in Table 4.16.

Table 4.16 *Descriptive Statistics of the Seven Predicative Indices for the Integrated Essay Scores: Training Set*

Coh-Metrix indices	Categories	M	S.D.	N
Number of words per text	Basic text information	197.220	52.222	160
Past participle verbs	Basic text information	26.920	16.301	160
Word familiarity (content words)	Lexical sophistication	569.710	4.851	160
Verbs in 3rd person singular present form	Basic text information	48.003	19.563	160
Semantic similarity (LSA sentence to sentence)	Cohesion	0.273	0.094	160
Verbs in base form	Basic text information	29.053	16.098	160
Word frequency (content words)	Lexical sophistication	2.297	0.137	160

The model demonstrated that the seven significant indices together explained 58.4% of the variance in the evaluation of the 160 integrated essays in the training set (see Table 4.17 for additional information). Twelve indices were removed from the regression model as being non-significant. These indices were nominalizations, noun hypernymy values, verbs in non-3rd person singular present form, personal pronouns, number of modifiers per noun phrase, word concreteness (content words), number of sentences per text, noun overlap, gerund or present participle verbs, tense repetition, prepositional phrases, and verbs in past tense. Table 4.17 presents detailed information of the seven indices that were retained in the regression model. *t*-test information of the seven indices together with the amount of score variance explained is shown in Table 4.18.

Table 4.17 *Regression Analysis Findings to Predict the Integrated Essay Scores: Training Set*

Entry	Coh-Metrix index added	Correlation	R^2	*B*	B	*S.E.*
Entry 1	Number of words per text	0.513	0.264	0.009	0.378	0.001
Entry 2	Past participle verbs	0.647	0.419	0.021	0.258	0.005
Entry 3	Word familiarity (content words)	0.710	0.504	-0.055	-0.206	0.018
Entry 4	Verbs in 3rd person singular present form	0.738	0.545	0.009	0.133	0.004
Entry 5	Semantic similarity(LSA sentence to sentence)	0.747	0.559	2.015	0.146	0.757
Entry 6	Verbs in base form	0.756	0.572	-0.011	-0.136	0.005
Entry 7	Word frequency (content words)	0.764	0.584	-1.348	-0.142	0.651

Notes: *B* =unstandardized β; B= standardized; S.E. = standard error. Estimated constant term is34.580.

Table 4.18 *t-value, p-values, and Variance Explained of the Seven Significant Indices for the Integrated Essay Scores: Training Set*

Coh-Metrix indices	*t*-value	*p*-value	R^2
Number of words per text	6.964	<0.001	0.264
Past participle verbs	4.176	<0.001	0.156
Word familiarity (content words)	-3.080	<0.050	0.085
Verbs in 3rd person singular present form	2.193	<0.050	0.041
Semantic similarity (LSA sentence to sentence)	2.662	<0.050	0.014
Verbs in base form	-2.081	<0.050	0.013
Word frequency (content words)	-2.071	<0.050	0.012

Test Set

Following Crossley and McNamara (in press, a), a test set (80 randomly selected integrated essays) was used to further validate the results yielded in the regression model based on the training set (160 integrated essays). To determine the predicting power of the significant predictors identified, an estimated score for each integrated essay in the independent test set (80 essays) was generated using the B weights and the constant from the training set regression analysis. I then conducted a Pearson's correlation between the estimated score and the actual score assigned on each of the integrated essays in the test set. This correlation together with its $r^2$was then calculated to determine the predictive accuracy of the training set regression model on the independent data set.

The model, when applied to the test set, produced r= .730, r^2 = .533. The results from the test set model thus demonstrated that the combination of the seven predictors accounted for 53.3% of the variance in the assigned scores of the 80 integrated essays in the test set.

Independent Essays

Correlation analysis demonstrated that the scores of the independent essay in the training set were significantly correlated with the following Coh-Metrix indices: 1) basic text information (number of sentences number of paragraphs, and number of words per text, verbs in base form

and in non-3rd person singular present form, past participle verbs, verb phrases, personal pronoun possessive cases), 2) lexical sophistication (average syllables per word, lexical diversity, word frequency, hypernymy and polysemy values, word concreteness, familiarity, imagability, and meaningfulness, and nominalizations), 3) syntactic complexity (number of modifiers per noun phrase), and 4) cohesion (conditional connectives, word overlap, and aspect repetition). After controlling for multi-collinearity, 21 indices met the requirement to enter the regression analysis. Results from the reported correlations are provided in Table 4.19 in the order of the strength of the correlation.

Table 4.19 *Selected Coh-Metrix Indices for Regression Analysis of the Independent Essays: Training Set*

Coh-Metrix indices	Categories	*r* value	*p* value
Number of words per text	Basic text information	.691	p<0.001
Nominalizations	Lexical sophistication	.521	p<0.001
Noun hypernymy values	Lexical sophistication	.475	p<0.001
Past participle verbs	Basic text information	.464	p<0.001
Verbs in non-3rd person singular present form	Basic text information	-.441	p<0.001
Word familiarity (all words)	Lexical sophistication	-.419	p<0.001
Lexical diversity D	Lexical sophistication	.415	p<0.001
Word meaningfulness (all words)	Lexical sophistication	-.365	p<0.001
Embedded clauses	Syntactic complexity	-.339	p<0.001
Number of modifiers per noun phrase	Syntactic complexity	.337	p<0.001
Average syllables per word	Lexical sophistication	.309	p<0.001
Aspect repetition	Cohesion	-.308	p<0.001
Personal pronouns	Basic text information	-.297	p<0.001
Word frequency (content words)	Lexical sophistication	-.295	p<0.001
Content word overlap	Cohesion	-.289	p<0.001
Verbs in base form	Basic text information	-.281	p<0.001
Conditionals connectives	Cohesion	-.245	P<0.050
Number of paragraphs per text	Basic text information	.209	P<0.050
Word polysemy values	Lexical sophistication	-.170	P<0.050
Word concreteness (content words)	Lexical sophistication	.167	P<0.050
Word imagability (all words)	Lexical sophistication	-.156	P<0.050

A regression analysis using the 21 selected indices to account for the variance in the essay scores was conducted for the training set of 160 independent essays. The regression

yielded a significant model, F (1, 154) = 57.325, p < .001, r = .807, r^2 = .650, adjusted r^2 = .639. Five Coh-Metrix indices were significant predictors in the regression: number of words per text, average syllables per word, noun hypernymy values, past participle verbs, and conditional connectives. Descriptive statistics for these five indices are provided in Table 4.20.

Table 4.20 *Descriptive Statistics of the Five Predicative Indices for the Independent Essay Scores: Training Set*

Coh-Metrix indices	Categories	M	S.D.	N
Number of words per text	Basic text information	309.060	77.543	160
Average syllables per word	Lexical sophistication	1.563	0.108	160
Noun hypernymy values	Lexical sophistication	5.485	0.508	160
Past participle verbs	Basic text information	13.557	9.691	160
Conditional connectives	Cohesion	4.164	4.776	160

The model demonstrated that the combination of the five variables accounted for 65.0% of the variance in the evaluation of the training set of 160 randomly selected independent essays (for additional information see Table 4.21). Table 4.22 presents *t*-test information of the five indices that were retained in the regression model and the score variance that each of them explained.

Table 4.21 *Regression Analysis Findings to Predict the Scores of Independent Essays: Training Set*

Entry	Coh-Metrix Index added	Correlation	R^2	*B*	B	*S.E.*
Entry 1	Number of words per text	0.691	0.478	0.007	0.577	0.001
Entry 2	Average syllables per word	0.753	0.568	1.511	0.179	0.448
Entry 3	Noun hypernymy values	0.785	0.616	0.359	0.199	0.094
Entry 4	Past participle verbs	0.800	0.641	0.016	0.165	0.005
Entry 5	Conditional connectives	0.807	0.650	-0.020	-0.104	0.010

Notes: *B* = unstandardized β; B = standardized; S.E. = standard error. Estimated constant term is -3.097.

Table 4. 22. *t-value, p-values, and Variance Explained of the Five Significant Indices for the Independent Essay Scores: Training Set*

Coh-Metrix indices	*t*-value	*p*-value	R^2
Number of words per text	11.194	<0.001	0.478
Average syllables per word	3.376	<0.050	0.090
Noun hypernymy values	3.837	<0.001	0.048

Past participle verbs	2.992	<0.050	0.025
Conditional connectives	-2.071	<0.050	0.010

Similar to the data set of the integrated essays, the results from the regression conducted on the training set of the independent essays were also extended to the test set (80 independent essays), which were withheld from the original analysis. Following the same procedure (using the B weights and the constant from the training set regression model to derive estimated scores for the essays in the test set), Pearson correlation between the estimated scores and the actual scores was calculated to estimate the predicting accuracy of the model on the test set.

The regression model, when extended to the test set of the independent essays, yielded $r = .758$, $r^2 = .574$, demonstrating that the combination of the five significant predictors identified in the training set regression model accounted for 57.4 % of the variance in the actual scores assigned on the 80 independent essays in the test set.

Discussion for Research Question 2

The regression analysis provided empirical evidence to illustrate that linguistic features can significantly predict evaluations of writing quality for the integrated and the independent essays. The analyses also demonstrated that the models established on the training sets can be extended to the independent data sets (test sets) and achieve similar predicting accuracy. The more rigorous statistical methodology adopted better controls for issues such as over-fitting (Crossley & McNamara, in press a). Thus, the study, lends reliable support to the theoretical argument that linguistic features vary with essay scores for both of the writing tasks. Furthermore, the findings also help to validate the scoring rubrics used for the two tasks by showing whether the descriptors detailed in the scoring rubric correspond with the significant predictors identified in the regression models. The following table summarizes the significant linguistic predictors for the integrated and the independent essay quality respectively.

Table 4.23 *Significant Predictors for Integrated and Independent Essay Scores*

Coh-Metrix indices	Integrated	Independent
Number of words per text	Yes	Yes
Past participle verbs	Yes	Yes
Word familiarity (content words)	Yes	No
Verbs in 3rd person singular present form	Yes	No
Semantic similarity (LSA sentence to sentence)	Yes	No
Verbs in base form	Yes	No
Word frequency (content words)	Yes	No
Average syllables per word	No	Yes
Noun hypernymy values	No	Yes
Conditional connectives	No	Yes

To fully explain, the following section will discuss the regression analysis results for the integrated essays followed by those for the independent essays.

Integrated Essays

Similar to previous studies on integrated writing (Gebril & Plakans, 2009; Watanabe, 2001), this study also demonstrated that textual length has a large effect on the essay scores assigned (defined as Pearson's correlations ≥ .50, Cohen, 1988). Longer essays were scored higher. In fact, as shown in Table 4.18, text length has the largest effect size among all the seven indices that were retained in the regression model and by itself accounts for 26.4% of the score variance of the integrated essays alone. This relationship between textual length and the essay scores is not difficult to understand as many of the features of highly scored essays (e.g., details to support a statement) are difficult to embed in a short essay (Chodorow & Burstein, 2004). Textual length being a strong predictor of essay scores has also been repeatedly verified in independent writing tasks (Chodorow & Burstein, 2004; Ferris, 1994; Frase et al., 1999; Reid, 1990).

The finding on past participle verbs adheres to expectations premised on research on independent essays. The study showed that the higher rated integrated essays contained more

occurrences of past participle verbs. Past participle verbs are normally used to construct passive voice or to indicate present or past aspect. A closer examination of the essays revealed that the past participle verbs in the integrated writing mainly occurred in construction of passive voice. Examples from the integrated essays include:

> The professor supports that species of fish that are **used** to feed the farm-raised fish are usually not **eaten** by people. (Independent20073293)
>
> Humans are "**exposed** to harmful or unnatural long-term effects" when consuming farm-raised fish, which are **fed** with growth-inducing chemicals. (Independent20073224)

Since passive voice is one of the markers for formal academic writing style (Hinkel, 2002), this finding suggests that the higher rated integrated essays include more linguistic devices that are characteristic of general academic writing. Although not included in the final regression model, the significant positive correlation between nominalizations and the integrated essay scores (see Table 4.15) also confirmed that the higher rated integrated essays bore more resemblance to formal academic writing than the lower rated ones. This particular finding has actually been reported in previous studies on independent writing as well: independent essays that are rated as higher quality include more instances of nominalizations and passive voice (Connor, 1990; Ferris, 1994; Grant & Ginther, 2000).

The findings about the integrated essays also demonstrated the potential for cohesion (as evidenced by semantic similarity (LSA sentence to sentence)) to predict the integrated essay scores. Comparing the scoring rubric of the integrated writing and that of the independent writing (Appendix A), it can be seen that the evaluation criteria principally focus on accurate and coherent presentation of the extracted information in the integrated essays in addition to grammatical accuracy. Meanwhile, in the independent scoring rubric, linguistic sophistication at lexical and syntactic levels is emphasized in addition to the logic and coherence of the arguments

and grammatical accuracy. To further illustrate, the rubric descriptors for the highest scores (5 points) for both the integrated and independent writing are presented below.

> Integrated writing scoring criteria (5 points)
>
> A response at this level successfully selects the important information from the lecture and coherently and accurately presents the information in relation to the relevant information presented in the reading. The response is well organized, and occasional language errors that are present do not result in inaccurate or imprecise presentation of content or connections.
>
> Independent writing scoring criteria (5 points)
>
> An essay at this level largely accomplishes all of the following:
> - Effectively addresses the topic and task
> - Is well organized and well developed, using clearly appropriate explanations, exemplifications, and/or details
> - Display unity, progression, and coherence
> - Displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice, and idiomaticity, though it may have minor lexical or grammatical errors.

The semantic similarity index indicates conceptual similarity between a sentence and every other sentence in a given text. The analysis showed that the higher rated integrated essays had a higher conceptual similarity than the essays that were judged to be of a poorer quality. Although not included in the regression analysis, many other cohesive devices also demonstrated a similar trend in the correlation analysis. For example, the higher rated integrated essays contained significantly more lexical overlap and aspect repetition than those rated of lower quality.

Additionally, the findings also revealed that the higher rated integrated essays contained significantly more verbs in 3rd person singular present form and fewer verbs in base form. As mentioned in research question 1, the frequent use of 3rd person singular form is related to citing sources and staying on topic (in this case, staying on the topic of fish farming rather than focusing on farmers or consumers in general). This can be taken as a sign that the higher rated essays contained more occurrences of correctly marked verbs for citing the source and staying on

the topic of fish farming, thus conveying expected information in a detached way (without using first or second person pronouns).

Meanwhile, as for the predictor of verbs in base form, its negative *t* value suggests that the essays including more such cases were actually rated lower. The low rated integrated essays that scored high on this Coh-Metrix index were pulled out from the corpus of essays for further examination. It was found that the majority of the verbs in base form were actually grammatical errors because the test takers failed to correctly indicate the subject of the sentence or did not provide the correct suffixes for the verbs. The following example is provided to illustrate:

> …because the fishes from the fish farm aren't **produce** to release into the wild, but rather for commercial purposes.(Integrated20073075)

Therefore, this significant predictor of verbs in base form suggests that, in terms of verb forms, the integrated essays that contained more grammatical errors were rated lower. Although grammatical accuracy was not reported in the computational analysis through Coh-Metrix, the regression analysis does suggest that grammatical accuracy plays in role in the evaluation of the integrated writing, which is consistent with findings from Cumming et al. (2005, 2006) and Gebril and Plakans (2009).

The regression analysis also exhibited some mixed findings related to lexical sophistication and syntactic complexity given the guidelines detailed in the scoring rubric (see Appendix A). On one hand, none of the Coh-Metrix indices of lexical diversity or syntactic complexity was included in the regression model as a significant predictor of the integrated essay scores. Although contradictory to the findings reported in many studies on independent writing (Engber, 1995; Grant & Ginther, 2000; Reppen, 1994), these findings are in accordance with the integrated scoring rubric. On the other hand, some lexical features were found to be significantly predictive of the integrated essay scores even though no such features are detailed in the scoring

rubric. For instance, the study found that word familiarity is a significant predictor of the integrated essay scores, accounting for 8.5% of the score variance. As shown in Table 4.18, the test takers who used fewer familiar words were given higher scores, suggesting a positive correlation between lexical sophistication and essay scores, a finding often made in independent essay scores (Crossley & McNamara, in press a). This finding provides evidence that lexical choice made by the test takers has a significant positive correlation with the scores even though the scoring rubric does not address these lexical choices. The very last predictor identified in the regression model, word frequency, is another lexical feature that was found to be able to differentiate writing quality perceived. The results (shown in Table 4.18) illustrates that the test takers who were judged to be more proficient used more words that are less frequent than those whose essays were rated less favorably. Since less frequent words indicate higher lexical sophistication, this particular finding reinforces the idea that lexical sophistication is predictive of the scores of the integrated essays. Therefore, two lexical predictors of the essay scores were actually identified. Although they both suggest the positive impact of lexical sophistication on the essay scores, a relationship often observed between lexical features and writing quality, the correlations cannot be directly predicted solely based on the scoring rubric. The findings therefore illustrate the phenomenon articulated by Lumley (2005) that in rating, raters might attend to many features beyond what is included in the scoring rubric.

In general, it can be seen that linguistic features do vary with the score levels in the integrated writing. The regression analysis showed that in the integrated task, the writing quality was at least partially determined by whether the expected content is being presented (as evidenced by verbs in 3rd person singular present form) and whether that information is presented cohesively (as evidenced by semantic similarity) and with good grammar (in terms of verb

forms). What is unexpected based on the specifics detailed in the scoring rubric is that higher level of lexical sophistication appeared to be significantly correlated with higher scores. However, in line with previous empirical research on writing quality (e.g., Nation, 1988), this finding should not be surprising.

Independent Essays

In accordance with previous studies on the effect of text length, the regression analysis also illustrated that text length is a significant, powerful predictor of the independent essay scores, accounting for 47.8% of the variance (as shown in Table 4.21). Similarly, the results related to average syllables per word and noun hypernymy values also parallel the findings reported in previous studies on lexical properties in relation to writing quality (Crossley & McNamara, in press a; Frase et al., 1999). The findings from the current study are also consistent with the scoring rubric of the independent writing. The test takers who were classified as more proficient writers were found to use more words that display a high level of sophistication (as evidenced by being more sophisticated and being more specific and unambiguous) as compared to those who received lower scores.

Similar to the findings made in the integrated essays, past participle verbs were also found to be a significant predictor of the essay scores, being positively correlated with the independent essay scores. Therefore, the test takers who were judged to be more proficient produced significantly more cases of past participle verbs in comparison to those who were judged to be less proficient. As mentioned earlier, the more frequent use of past participle verbs in the test takers' essays was correlated with the use of passive voice, indicating that higher proficient writers employed more passive voice structures, a feature of general academic writing (Hinkel, 2002), in their essays than the writers who were judged to be less proficient.

An index on cohesion was also included in the regression model of the independent essay scores. Conditional connectives were found to be a significant predictor of the independent essay scores with a negative correlation (as seen in Table 4.19). Mixed results have been reported as to whether high proficient writers produce more cohesive devices in their writing when compared to their counterparts with lower proficiency in writing (Connor, 1990; Crossley & McNamara, in press a; Jin, 2001). The findings in this study showed that the test takers who were rated to be more proficient actually produced essays with fewer cohesive devices. Additional support can also be found in the correlation analysis (see Table 4.19). For instance, the essays composed by the more proficient test takers not only contained fewer conditional connectives but also had lower scores for two other cohesive devices: aspect repetition and content word overlap. Furthermore, the lexical diversity index also demonstrated that the more proficient test takers provided less lexical overlap, a similar finding reported in Crossley and McNamara (in press a),a study that also employed regression analysis to explore the predictive power of Coh-Metrix indices on the independent essay scores.

It should also be pointed out that none of the indices related to syntactic complexity was found to be a significant predictor of scores for the independent essays. Although incongruous with the guidelines provided in the scoring rubric of the independent writing, this finding was also reported in Crossley and McNamara (in press a).

Research Question 3

The third research question focuses on whether the linguistic features of the essays vary with the academic experience of the test takers within each task type. If integrated writing and independent writing tasks are integral to academic literacy activities, it is then reasonable to speculate that test takers with more academic experience are more likely to produce texts that

contain greater cohesion, lexical sophistication, syntactic complexity, and more features that are markers of general academic writing than those with less such experience. This speculation is premised on the fact that more academic experience often means more exposure to and practice of the target language in general and the academic writing activities in particular. Furthermore, if integrated writing tasks better resemble academic writing activities as compared to independent writing tasks, the effect of academic experience on the linguistic features should be even more pronounced in the integrated writing.

As previously mentioned, 48 out of the 240 test takers indicated that they took the test to get enrolled in undergraduate programs while 51 wanted to apply to graduate programs. Even though it is not possible to pinpoint the exact academic experience these participants have, it is reasonable to assume that those applying to graduate programs have had more academic experience at the tertiary level than those applying to undergraduate programs. Therefore, one way ANOVA was performed to compare the 48 test takers with the 51 test takers to see whether the linguistic features of their essays differed within each task type. Additionally, an independent *t*-test was conducted on their essay scores to further investigate the possible differences.

Given that linguistic features in relation to basic text information, lexical sophistication, syntactic complexity, and cohesion have been found to be correlated with writing proficiency, all of these Coh-Metrix indices (all together 70 indices) were included for the ANOVA.

Results for Research Question 3

In the following sections, the ANOVA results for the linguistic features and the *t*-test results for the scores were reported for each task type respectively.

Integrated Essays

Because multiple ANOVAs have been conducted on the same data with the alpha level being set as .05, the overall chance of a type one error would increase. To control for alpha inflation, the original alpha level .05 divided by the number of ANOVAs was used. After controlling alpha inflation, one-way ANOVA results (see Appendix N) illustrated that none of the Coh-Metrix variables demonstrated significant differences among the two groups of test takers. Independent *t*-test of the integrated essay scores likewise did not show a significant difference between the two groups of test takers (Table 4.24).

Table 4. 24 *t-test Results and Means (standard deviations) for the Essay Scores across Undergraduate and Graduate Applicants*

	t	df	p	Undergraduate M(SD)	Graduate M(SD)
Integrated essay scores	.60	97	.953	3.156(1.380)	3.196(1.484)
Independent essay scores	-.138	97	.890	3.521(.978)	3.510 (.863)

Independent Essays

Similar to the integrated essays, one-way ANOVA of the 70 Coh-Metrix indices across the two groups of test takers did not yield any significant differences across the two groups of test takers. The ANOVA results can be found in Appendix O. Independent *t*-test of the independent essay scores again demonstrated no significant difference between the two groups of test takers (see Table 4.24).

Discussion for Research Question 3

None of the linguistic features showed a significant difference between the graduate and the undergraduate applicants in both the integrated and the independent writing tasks. Similar to the findings made in the textual analysis, no significant difference was located in the scores across the two groups of test takers in both tasks. Therefore, the findings revealed that the

linguistic features under investigation and the perceived quality of the integrated and the independent essays do not vary with the academic experience of the subset of the test takers who self-reported their goals in taking the test.

However, it is worth drawing attention to the fact that data analysis for research question 3 was based on the self-reported data of the test takers (their purposes for taking the TOEFL iBT). The findings, therefore, should be interpreted with caution for two reasons. First of all, no information as to the test takers' actual academic experience was available in the data set provided by ETS. Second, many of the 240 test takers did not specify their purposes in taking the test. Whether the clarification of the unspecified cases would change the results of the analysis is unfortunately unclear.

Summary of Quantitative Textual Analysis

The quantitative textual analysis section focused on investigating whether and how linguistic features varied with task type, essay scores, and academic experience of test takers. DA was performed to determine the task type difference. The results illustrated that, regardless of the writing quality, the linguistic features of the TOEFL iBT essays collected in 2007, mainly lexical features, can predict essay membership with 100% accuracy. As mentioned earlier, supplementary analysis with the 2006 data set on different topics also demonstrated that overall, the two types of essays differ from each other in similar ways. These results suggest that the linguistic differences between the two tasks found in this study were not just restricted to the two prompts under investigation. Instead, the linguistic patterns identified are more indicative of the differences between the text-based integrated writing and the independent writing. This finding is significant as it lends evidence to the theoretical claim that the integrated and independent

writing tasks can elicit different linguistic performance, thus broadening representation of the underlying writing construct in the writing test (Cumming et al., 2005, 2006; Huff et al., 2008).

The second research question focused on whether linguistic feature varied with essay scores within each task. Examining the regression models and the significant predictors, it can be concluded that for both of the tasks, certain linguistic features can significantly predict the essay quality perceived. The findings thus confirm that the linguistic features of the integrated as well as the independent essays do vary with the essay scores.

Comparing the predictor indices of the integrated essays with those of the independent essays, there are many similarities. First of all, text length was the predictor that explained the majority of the score variance identified in the regression analysis. That is to say, regardless of the task type, longer essays were evaluated more favorably than shorter ones. Secondly, although syntactic features are specified in the independent scoring rubric, none of the Coh-Metrix indices related to syntactic complexity were found to be significant predictors in either task. Thirdly, in terms of the lexical sophistication, higher rated independent essays were found to contain more instances of longer but more specific words than the lower rated essays. Similarly, although unspecified in the scoring rubric, it was also found that the integrated essays with higher scores included more cases of unfamiliar and less frequent words than the ones with lower scores. Finally, for both tasks, past participle verbs were found to be a significant predictor with a positive correlation with the essay scores. This suggests that either in the integrated or the independent writing, use of passive voice, a feature of general academic English, was significantly related to the scores. Taken together, all the similarities identified indicate that in both tasks, the linguistic features related to lexical sophistication and passive voice do vary with the essay scores in keeping with theoretical expectations.

The regression results also exhibited two major differences between the integrated and the independent essays. For the first difference, the integrated essays that were rated higher had a significantly higher score in semantic similarity (LSA sentence to sentence), which represents conceptual similarities among sentences. Meanwhile, there was also a cohesive device (conditional connectives) that was found to be able to predict the essay scores in the independent writing, but that device was negatively correlated with the essay scores. These findings demonstrated that in the integrated writing, the higher rated essays contained more cohesive devices while in the independent writing, the higher rated essays contained fewer cohesive devices than those rated lower. The second difference is that verbs in base form were a significant predictor of the integrated essay scores but not for the independent essay scores. The negative correlation indicated that higher rated integrated essays contained significantly fewer instances of verbs in base form than the lower rated essays. As mentioned earlier, verbs in base form were often grammatical mistakes (not correctly marked verb forms). Therefore, it seems that whether verb forms were correctly marked, one of the indicators of general language proficiency (Dulay, Burt, & Krashen, 1982; Ellis, 1994) had a significant correlation with the integrated essay scores but not with the independent essay scores. Although in both scoring rubrics, grammatical accuracy is listed as one of the descriptors, the results indicated that it had different relationship with the scores across the two tasks. Particular reasons for this difference are not clear. However, it should be noticed that although the index of verbs in base forms indicated grammatical accuracy of the essays to a degree, Coh-Metrix cannot directly report or measure grammatical mistakes like other computational tools such as *e-rater*. Additional information about grammatical accuracy (such as subject-verb agreement and incorrect word

forms) that can be directly reported by *e-rater* (Quilin, Higgins, & Wolff, 2009) is certainly needed to further investigate this difference between the two task types.

To answer the third research question about whether linguistic features vary with academic experience of test takers, the study showed that none of the Coh-Metrix indices demonstrated significant differences across the graduate and undergraduate applicants. Academic experience at the tertiary level does not seem to leave a noticeable trace in the linguistic choices made by the test takers while constructing the integrated as well as the independent essays. Interestingly, like the independent essay scores, the integrated writing scores did not illustrate a significant difference between the graduate and the undergraduate applicants although previous studies have reported such differences (Delaney, 2008; Trites & McGoarty, 2005). The reason, as previously mentioned, might be related to the reliance on self-reported data of academic experience in the current study.

Therefore, the quantitative textual analysis indicated that the linguistic features of the TOEFL iBT essays collected in 2007 varied with task type and essay scores. However, when it comes to the academic experience of the test takers, the study found that the linguistic features under investigation did not vary with this variable. How these findings contribute to the validity argument of the two tasks will be discussed together with the findings made in the writing process analysis component in the final chapter of the dissertation.

It should also be pointed out that compared to previous studies on the linguistic differences between integrated and independent writing (e.g., Cumming et al., 2005, 2006; Gebril & Plakans, 2009), this study focused on a text-based integrated writing task, a task type that is in great need of empirical evidence to shed light on the construct it taps into. Second, not only surface level but also deep level linguistic features that contribute to textual cohesion were

explored in the current study so that a more comprehensive understanding of the differences across the two types of writing was made possible.

Finally, it should be noted that in many cases, a direct comparison of the differences identified in the current study with those reported in the previous related studies is not feasible. First of all, the overwhelming majority of the previous studies comparing linguistic features between integrated and independent writing have focused on thematically-related integrated writing (Cumming et al., 2005, 2006; Gebril & Plakans, 2009) while little has been reported on text-based integrated writing features. Second, in the previous studies, when linguistic features (such as lexical sophistication) were examined, they were measured very differently from the current study. For instance, lexical sophistication was measured by TTR in Cumming et al. (2005, 2006) while in this study lexical sophistication was measured through many different approaches such as lexical diversity, lexical frequency, word hypernymy values. Furthermore, not only different measures were employed to investigate individual linguistic features but also these measures were statistically more valid and reliable than traditional ones. For instance, as mentioned previously, different from Cumming et al. (2005, 2006), lexical diversity was measured through MTLD and D rather than TTR which is highly correlated with text length.

CHAPTER 5

QUALITATIVE PROCESS ANALYSIS

This portion of the dissertation looks at the results of the process analysis component of the study. The chapter begins with an introduction of demographic information of the 20 participants followed by a summary of the information collected from the post-task questionnaires and interviews. Then descriptive information of the writing behaviors that emerged from the participants' TAP data is provided. The subsequent sections present the results and discussions for the research questions 4, 5, and 6 respectively.

Demographic Information of the Participants

In this section, detailed information about the participants is presented. A total of 20 participants (10 undergraduate and 10 graduate students) participated in the think-aloud writing sessions. As mentioned in Chapter 3, no more than three participants were from the same linguistic or disciplinary backgrounds at either the graduate or the undergraduate level. Basic information about these participants is provided in Table 5.1. Each participant was given a pseudonym for the sake of confidentiality.

Table 5.1 *Characteristics of the Participants*

Participants	Gender	Academic status	Home country	Native language	Majors	TOEFL writing scores[a]
Jane	female	Undergraduate	Israel	Hebrew	Modern language	15
Kris	male	Undergraduate	Ivory Coast	French	Math	25
Victoria	female	Undergraduate	Vietnam	Vietnamese	Asian studies	20*
Elaine	female	Undergraduate	France	French	Hospitality	19
Ted	male	Undergraduate	China	Chinese	Chemistry	20
Julia	female	Undergraduate	Italy	Italian	Economics	23
Kevin	male	Undergraduate	Nigeria	Yoruba	Biology	20
Henry	male	Undergraduate	Brazil	Portuguese	Finance	26*
Larry	male	Undergraduate	Georgia	Georgian	Math	20
Sam	male	Undergraduate	India	Guajarati	Exercise education	20
Patrick	male	Graduate	Russia	Russian	Finance	20
Tina	female	Graduate	China	Chinese	Language education	22
Karren	female	Graduate	Japan	Japanese	Social work	18
Mark	male	Graduate	Korea	Korean	Public management	25

Kathy	female	Graduate	Nepal	Nepali	Computer science	25*
Aaron	male	Graduate	Nepal	Nepali	Chemistry	24
Gloria	female	Graduate	China	Chinese	Managerial science	24
Lora	female	Graduate	China	Chinese	Chemistry	22
Luke	male	Graduate	Brazil	Portuguese	Journalism	20
Mary	female	Graduate	India	Indian	Biology	24

[a] Self-reported scores from the most recent TOEFL iBT.
* They took computer based or paper based TOEFL instead of TOEFL iBT. The scores listed are already converted to the TOEFL iBT scores.

As can be seen, among the 20 participants, 10 of them were female, and 10 of them were male. At either the undergraduate or the graduate level, the participants represented both physical and social sciences. According to the self-evaluation of their writing ability in English (as revealed in the questionnaires), 50% of the participants reported that they were good writers in English, 40% were not sure, and 10% of them did not think that they were good writers. Meanwhile, 14 out of the 20 participants reported that they enjoyed writing in English, and 6 of them were not sure about it. Each of the participants had taken English writing courses either in their home country or here in the United States. All of them had experiences with summary writing in addition to expository and argumentative writing. All the participants had taken the TOEFL. Except for three of the participants (as shown in Table 5.1), everyone took the TOEFL iBT. The three participants' writing scores were converted to TOEFL iBT writing scores based on the criteria provided by ETS to allow for comparison. The mean writing score of the participants was 21.6 points with the minimum being 15 points and the maximum being 26 points.

Information Collected in Post-task Questionnaires and Interviews

According to the information revealed in the post-task questionnaires and interviews, all the participants reported treating the writing tasks as if they were real tests. Because of the articulation processes, they all took longer to compose their texts than what the tasks specified in normal testing conditions. For the integrated writing task, the average time to complete the task

was 42 minutes (ranged from 29 to 55 minutes), and for the independent writing task, the average time was 50 minutes (ranged from 43 to 78 minutes). None of the participants mentioned that thinking aloud altered their writing or thinking processes, but they all admitted that articulation slowed them down, especially for idea generation.

The participants were split in their views on the comparative difficulty of the two tasks. Twelve of the participants mentioned that the integrated writing task was easier because it was less cognitive demanding with the content being provided. Seven participants expressed the opposite opinion, reporting that the independent writing task was easier mainly because they were able to express and organize their thoughts without being constrained by any given materials. Only one participant indicated that the two tasks were of the same degree of difficulty.

The interview data also illustrated that the participants tended to have a shared understanding of the expected format in the independent writing. They all understood that they were expected to write an argumentative essay with an introduction, multiple body paragraphs, and a conclusion, although one of the participants deliberately chose not to write a conclusion. However, when it came to the integrated writing, the participants seemed to differ greatly in their interpretation of the expected format. Some of them held the opinion that they were also supposed to compose an essay with an introduction, body paragraphs, and a conclusion. Some of them thought that they were expected to write a one-paragraph summary, while others believed that their response should have a structure similar to that of the reading passage (an introduction paragraph in addition to three body paragraphs).

Information about the TAP Data

All the TAP data reported by the 20 participants was transcribed verbatim. The coding scheme that was used to segment the TAP data into individual writing behaviors is presented in

Table 5.2. Some of the writing behaviors were only found in one or the other task. When necessary, examples are also provided in the table to further illustrate the identified writing behaviors.

Table 5.2 *Coding Scheme for the TAP Data*

Categories	Definition	Examples
Commenting on one's understanding of source texts	Commenting on one's own understanding of the source texts	*Third point is the one I have more information, obviously, because I just understood better.* (Elaine-integrated)
Commenting on one's writing process	Commenting on the procedures taken to complete the writing task	*I will move on to the second point, and I will correct the first paragraph after.* (Julia-independent)
Commenting on one's writing product	Summarizing and evaluating what has been written and explaining why it has been written	*I should not repeat use "endanger." I should use another word.* (Ted-integrated)
Commenting on relationship between source texts	Comparing and contrasting the ideas presented in the source texts	*So in the reading passage I could be, I could see some counter arguments about, about…So they are against the, the fish farming. The opposite, in the, in the lecture, I could see some refute for this argument.* (Luke)
Global planning	Generating ideas; planning on how to organize the essay or the paragraph (attention is directed toward planning at the essay or paragraph level)	*So, I should start introducing, start introducing the two points first, or, I should summarize the reading part first? Yeah, I should start introducing the reading passage.* (Kris-integrated)
Planning and rehearsal	Developing local plans on what to say next and/or rehearsing different versions of wording and phrasing	*Collaborate? What the difference between collaborate and cooperate?* (Victoria-independent)
Positioning self	Choosing one's own stance on the given topic; considering pros and cons of different options and evaluating them	*So I say I agree that to cooperate well with each other is important, but it is not far more important than it was in the past.* (Gloria)
Reading one's writing	Reading what has been written	
Reading the instruction	Reading the task instruction	
Referring to notes	Reading one's own notes	What they say about in the lecture? [Start reading his notes] (Luke-integrated)

Referring to source texts	Reading and rereading source texts	Here, the reading passage say [then start reading the passage]. (Sam-integrated)
Revising and editing	Making an visible changes to what has been written	*(…more specialized works.) No, should be "jobs"* [Change "works" to "jobs"]. (Mark-independent)
Summarizing source texts	Summarizing what each source text is about; identifying the key ideas of source texts	*Ok, it's ["it" refers to the topic] about the fish farming.* (Mark-integrated)
Analyzing the task	Summarizing the task in one's own words; reviewing task requirement such as length, topic, structure, etc	*So they want me to write about the difference between the cooperation nowadays with the cooperation in the past.* (Henry-independent)
Unrelated comments	Comments that do not belong to any of the above categories	*Ok, environment. I hate this word (Jane-independent)*
Verbalizing one's writing	Verbalizing what is being written	*Fish farming has has positive po-si-tive effects.* (Victoria-integrated)

For the category of revising and editing, the corresponding writing behaviors were further divided into global revising and editing and local revising and editing following the guidelines proposed by Worden (2009). Local revising and editing refers to revisions made within sentence boundaries, and these types of changes usually do not affect global information presentation and understanding in the essay. On the other hand, global revising and editing is the changes that affect more than one sentence at a time. Generally speaking, global revising and editing tends to have a broader impact on the meaning making than the local revising and editing.

Table 5.3 summarizes the results that feature how the 20 participants used various writing behaviors in responding to the integrated and the independent writing tasks. The writing behaviors are presented in the order of frequency for the two tasks combined.

Table 5.3 *Number and Percentage of the Participants' Writing Behaviors*

	Total		Integrated		Independent	
Writing behaviors	Number	%	Number	%	Number	%
Verbalizing one's writing	1560	28.40	677	28.17	883	28.58
Revising and editing	1162	21.15	440	18.31	722	23.37
Planning and rehearsal	932	17.00	410	17.06	522	16.89
Reading one's writing	678	12.34	248	10.32	430	13.91
Commenting on one's writing product	395	7.19	198	8.24	197	6.38
Global planning	194	3.53	48	2.00	146	4.72
Reading the instruction	127	2.31	33	1.37	94	3.04
Referring to notes	104	1.89	86	3.58	18	0.58
Commenting on one's writing process	65	1.18	36	1.50	29	0.94
Analyzing the task	36	0.66	12	0.50	24	0.78
Unrelated comments	5	0.09	1	0.04	4	0.13
Summarizing source texts	94	1.71	94	3.91	0	0
Referring to source texts	87	1.58	87	3.62	0	0
Commenting on relationship between source texts	11	0.20	11	0.46	0	0
Commenting on understanding of source texts	21	0.38	21	0.87	0	0
Positioning self	21	0.38	0	0	21	0.68
Total	5493		2403		3090	

All together, 40 protocols containing 5493 writing behaviors were analyzed. As the table illustrates, 10 of the 16 behaviors were infrequently used and had an average below 2.5% of all the behaviors reported by the participants. The most frequently observed writing behavior is "verbalizing one's writing" (M = 28.4%). The frequent use of this behavior can be taken as an artifact of think-aloud writing sessions (Yang & Shi, 2003). The next most frequently used behavior was "revising and editing" (M = 21.15%), followed by "planning and rehearsal" (M = 17%), "reading one's writing" (M = 12.34%), "commenting on one's writing product" (M = 7.19%), and "global planning" (M = 3.53%). In the following sections, these protocols were examined for the task type, the essay scores, and the academic experience of the participants

respectively. The participants' responses to the demographic questionnaire, the post-task questionnaires, and the interviews were also incorporated to help interpret the patterns illustrated in the TAP data.

Research Question 4

Research question 4 concerns whether the writing behaviors that the participants employed varied across the integrated and the independent writing tasks.

Results for Research Question 4

Comparing the writing behaviors adopted by the 20 participants in responding to the two writing tasks, the very first similarity to be noticed is that for each of the tasks, the participants were involved in cyclical processes of planning, drafting, revising and editing. Each participant generated a non-linear writing process in completing the integrated and the independent tasks. This recursive pattern is illustrated in the following excerpts from the TAP data of Kris, an undergraduate participant. The first excerpt presented in Table 5.4 is from his integrated writing, and the second excerpt presented in Table 5.5 is from his independent writing.

Table 5.4 *Recursive Writing Behaviors in the Integrated Task*

Coding	TAP data	Writing behaviors
U02028*	And fish farming actually, actually, what? They helped the commercial fishing	Planning and rehearsal
U02029	And fish farming actually	Verbalizing one's writing
U02030	Um, no, [Delete "Fish farming actually"]	Revising and editing (local)
U02031	I should use in addition, in addition	Planning and rehearsal
U02032	In addition, fish farming was already	Verbalizing one's writing
U02033	No, no, no, [Delete "was already"]	Revising and editing (local)
U02034	They, they actually help the local population with commercial fishing	Planning and rehearsal
U02035	helps the local population with commercial fishing	verbalizing one's writing

*U02028 indicates that this piece of data is the 28th writing behavior reported by the second undergraduate participant.

Table 5.5 *Recursive Writing Behaviors in the Independent Task*

Coding	TAP data	Writing behaviors
U02189	If it was not, if there was no, if there was not, if the technology was a secret, if the, if this technology	Planning and rehearsal
U02190	If, if, this technology	Verbalizing one's writing
U02191	No device. [Change "technology" to "device"]	Revising and editing (local)
U02192	I already say that sentence, say that sentence.	Commenting on one's writing product
U02193	Uh [Delete "If this device"]	Revising and editing (local)
U02194	It shows that, yeah, it shows we can use them everywhere	Planning and rehearsal
U02195	It shows that we can use the technology everywhere in the world	Verbalizing one's writing
U02196	With any service [Add "with any service" before "we can"]	Revising and editing (local)

Before a comparison of the specific categories of the writing behaviors, it should be noted that the independent writing tasks generated significantly more writing behaviors than the integrated writing task as revealed in the independent *t*-test. This result is not difficult to understand given that the independent essays are significantly longer than the integrated essay. The *t*-test results for the number of writing behaviors and essay length across the two tasks are presented in Table 5.6. For this reason, comparison of the integrated and the independent writing tasks was made on basis of percentages rather than on pure counts of the writing behaviors in the subsequent analysis (Durst, 1987).

Table 5.6 *t-test Results and Means (standard deviations) for the Number of Writing Behaviors and Essay Length*

	t	df	p	Independent M(SD)	Integrated M(SD)
Number of writing behaviors	3.121	19	<.01	154.45 (51.84)	120.15 (41.00)
Essay length	4.896	19	<.01	329.70 (66.56)	231.60 (65.11)

To investigate the similarities and differences in the reported writing behaviors across the two tasks, each of the categories of the writing behaviors was examined. Figure 5.1 illustrates all

the categories of the writing behaviors (percentage) generated by the integrated and the independent writing respectively. The writing behaviors again are listed in the order of their percentage in the two tasks combined.

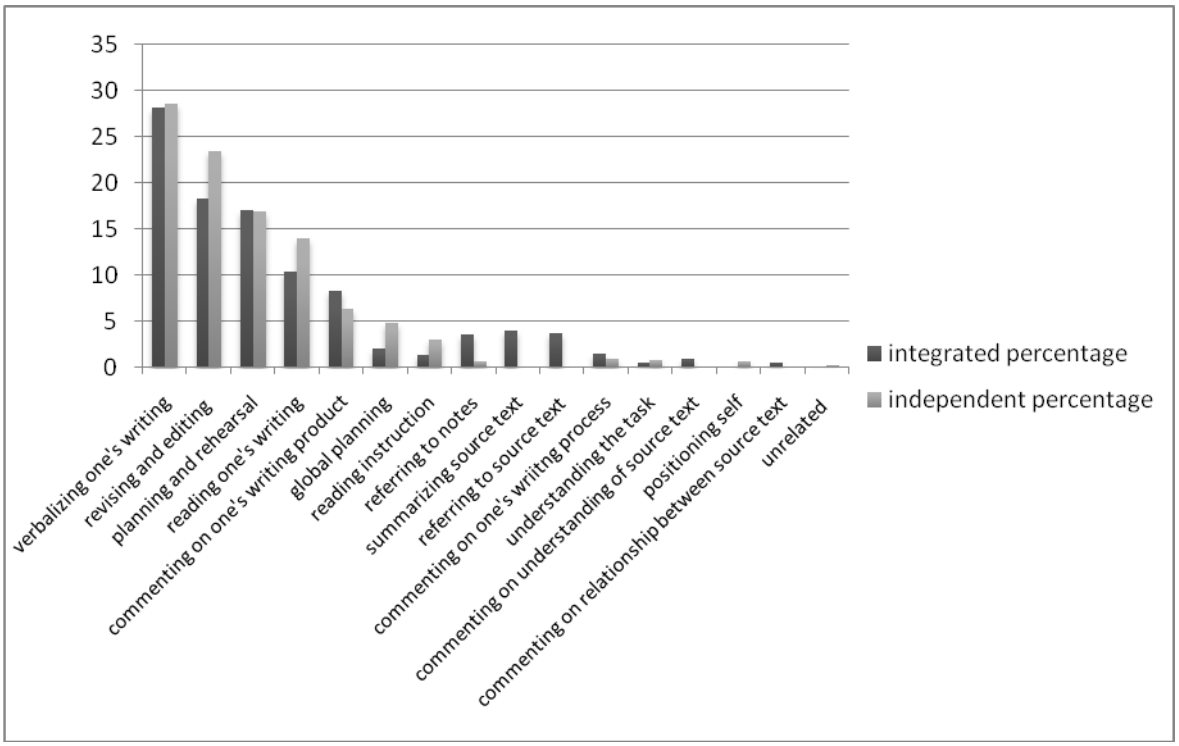


Figure 5.1 *Percentage of Each of the Writing Behaviors across the Two Writing Tasks*

The figure above shows that the integrated and the independent writing tasks shared many of the writing behaviors. We can also see in Table 5.3 that actually 11 out of the 16 categories of the writing behaviors occurred in both types of writing. As demonstrated inFigure 5.1, the shared writing behaviors followed almost the same order of percentage across the two tasks. For both tasks, the four categories of "verbalizing one's writing," "revising and editing," "planning and rehearsal," and "reading one's writing" occurred most frequently, and together they accounted for more than 70% of all the writing behaviors (see Table 5.3). Furthermore, after controlling for alpha inflation, a Wilcoxon Signed-ranks test results indicated that, among the shared categories of the writing behaviors, the 20 participants did not differ significantly across

the two tasks in the following eight categories: “verbalizing one’s writing,” “revising and editing,” “planning and rehearsal,” “reading one’s writing,” “commenting on one’s writing product,” “commenting on one’s writing process,” “analyzing the task,” and unrelated comments.

As for the other three shared categories of the writing behaviors, there are significant differences in the percentage of their occurrences across the two tasks. The Wilcoxon Signed-ranks test results of the three categories are presented in Table 5.7.

Table 5.7 *Wilcoxon Signed-ranks Test Results for the Three Categories of Shared Writing Behaviors*

	Integrated Mdn	Independent Mdn	*Z*	*p*	*r*
Referring to notes	3.17	0.52	3.845	0.000	0.608
Reading the instruction	1.61	3.62	3.041	0.002	0.481
Global planning	1.95	4.43	2.987	0.003	0.472

As shown in Table 5.7, as far as the 20 participants are concerned, the writing behavior of “referring to notes” occurred significantly more frequently in the integrated writing task than in the independent writing task. In the integrated writing, the participants referred to their notes to retrieve the writing content provided in the source texts. The majority of the participants not only took notes on the listening passage but on the reading passage as well. In the independent writing, many of the participants also took notes before they started composing their essays. These notes very often were general outlines for their independent essay writing. During the process of composing the independent essays, the participants also referred to their notes to recall the overarching plans they made a priori. However, this behavior occurred at a significantly lower percentage in the independent writing as compared to that in the integrated writing (0.58% vs. 3.58%). On the other hand, the 20 participants used “reading the instruction” and “global planning” at a significantly higher percentage in the independent writing task than in the integrated writing task.

In addition to the differences in the shared categories of the writing behaviors, the integrated and the independent writing tasks were also different in that particular patterns were associated with each of the writing tasks. In the integrated writing (as shown in Table 5.3), due to the nature of composing from the source texts, the participants generated some writing behaviors that were unique to this particular writing task. These writing behaviors include, listed in the order of percentage, "summarizing source texts" (3.91%), "referring to source texts" (3.62%), "commenting on understanding of source texts" (0.87%), and "commenting on relationship between source texts" (0.46%). All together, these categories accounted for 8.86% of all the writing behaviors elicited by the integrated writing task. Although the majority of these writing behaviors occurred in the pre-writing stage, the participants also reported them during the process of composing their integrated essays. Meanwhile, the independent writing task had a category of "positioning self" (0.68%), which did not occur in the inventory of the writing behaviors of the integrated writing.

Discussion for Research Question 4

First of all, by examining the order in which the various writing behaviors occurred, both types of writing turned out to be similar. The participants were all involved in a series of non-linear processes of analyzing, planning, drafting, revising and editing while completing the two types of writing tasks. The finding that, in both the independent and the text-based integrated writing, the writing procedure was not linear but circular and recursive is in accordance with previous research on writing processes (Flower& Hayes, 1981; Plakans, 2007). This finding suggests that the integrated writing task as well as the independent writing task are valid in the sense that they both elicit a series of recursive writing behaviors similar to what writers engage

in non-testing academic setting (Murray, 1982), thus solidifying the connection between test tasks and the target language use (Bachman, 1990).

Secondly, similar to what Plakans (2007) found, the participants in this study were also engaged in many of the same types of writing behaviors when completing the independent and the text-based integrated writing tasks. For both tasks, the participants made great effort to monitor their text construction practice. They were constantly involved in planning their content, rehearsing different ways to phrase and word the ideas to be conveyed, and monitoring their writing product through rereading and revising and editing.

With this being said, it is important to draw attention to the fact that, although the content being provided might lead to verbatim source use, that did not occur with my participants. They still spent almost the same amount of effort on planning and rehearsal (generating detailed content and planning for word and phrase choices) as they did in the independent task. An example taken from Luke's (a graduate participant) TAP data is provided below in Table 5.8. The excerpt helps to illustrate that the participant did not completely rely on borrowing words or phrases directly from the source texts. Instead, even when the content was provided, the participant still spent time figuring out how to put the extracted information in his own words while responding to the integrated writing task.

Table 5.8 *Use of Planning and Rehearsal in the Integrated Writing*

Coding	TAP data	Writing behaviors
G09048	So let me start with "however"	planning and rehearsal
G09049	However, the lecture	verbalizing one's writing
G09050	Says, no, have, no, presents,	planning and rehearsal
G09051	The lecture presents different view about the issue. According to the lecture,	verbalizing one's writing
G09052	What they talk about in the lecture?	referring to notes
G09053	According to the lecture, it, the fish, is already in danger	planning and rehearsal

G09054	The wild fishes in the region are already endangered and because the	verbalizing one's writing
G09055	The wild fishes in the region are already endangered and because	reading one's writing
G09056	No, no, [Add "it is happening" before "because"]	revising and editing (local)
G09057	But because the, the hunting?	planning and rehearsal
G09058	Because the hunting, fish…	verbalizing one's writing

Thirdly, examination of the particular categories of the writing behaviors associated with each of the writing tasks indicates that the participants, while composing the integrated essays, interacted frequently with the reading text (by constantly referring back to the reading text) as well as the notes they created from the extracted information out of the source texts. All the participants interacted with the source texts both before and during composing. As demonstrated in the TAP data and the post-task questionnaires and interviews, in preparing to write, almost all the participants summarized the source texts and explicitly commented on the relationship between the two texts. Therefore, similar to previous studies on integrated writing processes (Esameli, 2002; Plakans, 2007), this analysis also demonstrated that the integrated writing elicited many inter-textual writing behaviors that connect the source texts and the essays that the participants produced. More specifically, the participants interacted with the source texts by summarizing the content of and the relationship between the two source texts, reviewing the reading passage and their notes, and commenting on their understanding of the source texts. Through these active and purposeful interaction with the source texts, the participants were able to (re)construct the meaning embedded in the source texts, single out important information, and synthesize the information in their own writing (Spivey, 1997). Therefore, instead of verbatim source use, the participants were actually found to be involved in meaningful interaction with the source texts in the reading/listening to write process. Following is an example taken from Julia's

(an undergraduate participant) TAP data to demonstrate how she interacted with the source texts before she started writing:

Table 5.9 *Interacting with the Source Texts in the Integrated Writing*

Coding	TAP data	Writing behaviors
U06005	So, so, [the reading passage] the introduction and three negative aspects of it [fish farming] and there's nothing else.	summarizing source text
U06006	So the readings go, the listening goes against the readings?	commenting on the relationship of source texts

According to some of the participants, interacting with the source texts not only helped them to retrieve the content but also enabled them to foresee the structure of the integrated essays to be constructed. As Luke commented in the semi-structured interview, summarizing the source texts allowed him to realize that "there are three points, so, so basically, three, three main paragraphs to write. That is clear." In this sense, the participants were involved in mining the source texts in a writerly way (Church & Bereiter, 1984; Greene, 1992); they processed the source texts not only to extract the content information but also to derive the rhetorical structure of their own writing. For this reason, many participants found it unnecessary to explicitly plan the number of paragraphs to compose and decide the main idea to cover in each of the paragraphs, the primary function of the category of global planning. This might explain why the participants spent significantly less time on global planning in the integrated writing task than they did in the independent writing task (see Table 5.3), a similar finding reported in Plakans (2007). Taken together, these behaviors strongly suggest that the participants adopted a writing process that was interactive with the source texts in the integrated writing task. Rather than using the information and language directly from the source texts, they were actually involved in reading/listening to learn (Grabe, 2001) or discourse synthesis (Plakans, 2008, 2009) rather than

superficial comprehension of the source texts. These interactive writing behaviors thus indicate that successful completion of the task required the participants to be actively engaged with the source texts.

In addition, some participants also mentioned that they were able to determine the accuracy and completeness of the information presented in their own essays by referring to the source texts and the notes. Jane, in the semi-structured interview, stated: "oh, I have to read my notes, read the passage. I want make sure, I include everything, everything I write down. The order is also important. I don't want to change the order." This may be one reason why, compared to the independent writing task, the participants referred to their notes significantly more frequently in the integrated writing task than in the independent writing task. Furthermore, in several cases of interacting with the source texts, the participants also evaluated their understanding of the source texts. For instance, Elaine commented on her understanding of the listening passage by saying, "I have more to say about the third point, because, because I just understood better." It is, however, should be noted that some of the participants reported that being aware of the contradictory viewpoints presented in the two source texts (as indicated in the instructions) allowed them to utilize the reading passage to check for the accuracy of their understanding of the listening passage.

Interestingly, although the interview data illustrated that the participants differed from each other in their understanding of the integrated writing task, they spent significantly more time reading the instruction and the prompt in the independent writing task than in the integrated writing task. It suggests that although the task interpretation was different, the participants seemed to be confident of the task interpretation and the expected format they themselves constructed in the integrated writing. On the contrary, in the independent writing, the task

required the participants to establish their own stance on the imposed topic which might be unfamiliar to them. Many of the participants very often read and reread the instructions and the prompt and spent more time figuring out what the question was asking them to do in terms of content. In the semi-structured interview, Karen, a graduate participant from Japan, articulated that "I reread the prompt, for, for checking my understanding. I want to see that I understand it right. Not write on a wrong topic."

Research Question 5

Research question 5 investigates whether the writing behaviors undertaken by the participants varied with the essay scores in the integrated writing as well as in the independent writing. To derive the scores, the 20 integrated and 20 independent essays were rated by two raters using the scoring rubrics provided by ETS. Both raters had extensive experience in rating ESL writing. To train the two raters, the rubrics were fully explained. The raters were also required to score 25 essays taken from the 2007 corpus of the integrated and independent essays to norm this grading. The two raters achieved matching or adjacent scores (score difference less than 1 point) for 85% of the essays in their initial rating. The final scores were the average of the two raters' scores. In the cases of score discrepancy, the differences between the raters were resolved through discussion for the final scores. The scores (although not provided by ETS-trained raters) served the purpose of dividing the participants into high or low performance groups. The 20 participants were divided into two groups within each task type: the high performance group (the participants who scored no less than 3.5 points) and the low performance group (the participants who scored lower than 3.5 points). The final scores of the 20 integrated and the 20 independent essays together with their grouping information are presented in Table 5.10 in the order of the integrated essay scores.

Table 5.10 *Final Scores and Group Information of the Integrated and Independent Essays*

		Integrated		Independent	
Participants	Academic status	Scores	High/Low groups	scores	High/low groups
Luke	graduate	4	High	5	High
Mary	graduate	4	High	3	Low
Julia	undergraduate	4	High	5	High
Gloria	graduate	4	High	4.5	High
Aaron	graduate	4	High	4	High
Kathy	graduate	4	High	4	High
Mark	graduate	4	High	4	High
Tina	graduate	4	High	4	High
Jane	undergraduate	3.5	High	4	High
Henry	undergraduate	3	Low	3	Low
Patrick	graduate	3	Low	3	Low
Ted	undergraduate	3	Low	3	Low
Sam	undergraduate	3	Low	4	High
Lora	graduate	3	Low	4	High
Victoria	undergraduate	3	Low	4	High
Karen	graduate	3	Low	4	High
Larry	undergraduate	2.5	Low	3	Low
Kris	undergraduate	2.5	Low	4	High
Kevin	undergraduate	2	Low	2.5	Low
Elaine	undergraduate	2	Low	3	Low

Detailed information about the number of participants at each score level together with the descriptive statistics of the scores is presented in Table 5.11.

Table 5.11 *Number of Participants at Each Score Level and Descriptive Statistics of the Scores*

Scores	Integrated	Independent
5	0	2
4-4.5	8	11
3-3.5	8	6
2-2.5	4	1
M	3.275	3.750
S.D.	0.697	0.698

Since the participants involved in the TAP writing sessions were all matriculated ESL students, their writing proficiency was already prescreened through university admission procedures. Therefore, it is not surprising to find that there were no essays with scores lower than 2 points and the average scores for either the integrated or the independent essays were above 3 points.

As can be seen from Table 5.10, there were all together 11 low performing participants and nine high performing participants in the integrated task while there were only seven low performing participants and 13 high performing participants in the independent writing task.

Results for Research Question 5

The results regarding how the writing behaviors reported by the high performance participants compared to those of the low performance participants in the integrated writing will be presented first. The corresponding results in the independent writing will be reviewed in the subsequent sections.

Integrated Writing Task

Comparing the total number of writing behaviors utilized by the high and the low performance participants, it was found that there were no significant differences between the two groups. Therefore, in the following analysis, the number of writing behaviors was used directly for comparison.

The results of the Mann-Whitney test revealed that, for all the categories of the writing behaviors that were elicited by the integrated writing task, none of them demonstrated a significant difference across the two groups of participants. Therefore, no significant writing behavior differences were identified, suggesting that the two groups of participants utilized similar writing behaviors in constructing their integrated essays.

Independent Writing Task

Similar to the integrated writing task, the two groups of participants did not differ significantly in terms of the total number of writing behaviors produced. Again, the number of writing behaviors in each category was used directly to identify possible differences across the high performance and the low performance groups. Mann-Whitney test results also demonstrated that there were no significant differences across the two groups of participants.

Discussion for Research Question 5

As can be seen from the Mann-Whitney test results, for both the integrated and the independent writing, the high performance participants were similar to their low performance counterparts in all the writing behaviors reported. Therefore, the results illustrated that the writing behaviors of the 20 participants did not vary with their essay scores. The finding that the writing behaviors did not change with essay scores in the integrated writing is incongruous with Yang (2009), which reported that the high performance test takers interacted with the source texts more extensively and critically than the low performance ones. One possible explanation is that in Yang (2009), a post-task checklist was used to elicit information about the writing behaviors. That data collected was retrospective in nature. However, in this study, TAP data was examined. TAP data is more immediate and presumably reveals more about what is actually experienced by writers compared to self-reported retrospective data (Ericsson & Simon, 1993). Therefore, the different findings might be related to the different methods of data collection. Another possible explanation is that the participants involved in this study were all matriculated students, while in Yang (2009) the participants had a wider range of writing proficiency. In the current study, all the ESL participants had taken English writing courses, and all had experience with integrated writing in English. Such experience might have enabled them to adopt similar writing behaviors although their control of the form and information presentation in English

might still vary. However, in Yang (2009), in addition to matriculated undergraduate and graduate participants, pre-matriculated students (students enrolled in the ESL program) participated.

Research Question 6

Research question 6 examines the use of the writing behaviors in relation to the academic experience of the participants. The main goal is to find out whether the writing behaviors varied with the academic experience of the participants within the integrated writing task and within the independent writing task, and if so, whether the changes followed the same pattern across the two tasks.

In the context of the qualitative process component of the study, the academic experience was operationalized as the academic status of the participants (graduate vs. undergraduate). Given that the graduate participants had already completed their undergraduate studies, it is reasonable to assume that they had more experience with academic writing than the undergraduate participants and thus were more familiar with the writing tasks commonly assigned in higher education.

Many scholars have argued that integrated writing bears more resemblance to writing tasks assigned in higher education than independent writing (Cumming et al., 2000; Cumming et al., 2005, 2006; Lewkowicz, 1997; Weigle, 2004). Writing process studies in both testing and non-testing situations have showed that writers with more related experience, as compared to less experienced writers, tend to be more focused on global planning and revising than on sentence level issues in general and more engaged with source texts and more in control of their comprehension of the source texts in integrated writing (Kennedy, 1985; Plakans, 2008; Taylor & Beach, 1984). For these reasons, we would expect similar differences in writing behaviors adopted by the graduate and undergraduate participants. If it is the case, we would also expect

that the difference should be more pronounced with the integrated writing task than with the independent writing task. Therefore, comparison of writing processes between the two groups of participants (graduate vs. undergraduate) was made not only within each writing task but also across the writing tasks.

To gain more insight into the effect of academic experience on the test performance, the essays scores were also investigated to see whether it is a confounding factor. In the following sections, the results of how the writing behaviors varied across the graduate and the undergraduate participants will be presented in the order of the integrated and the independent tasks.

Results for Research Question 6

Examination of the scores received by the graduate participants and the undergraduate participants revealed a slightly different picture across the two tasks. In the integrated writing, judging by the scores assigned, the graduate participants were found to significantly outperform the undergraduate participants ($t = 3.40$, df = 18, $p = .003$). In contrast, no significant score difference was found with the independent writing task ($t = 1.305$, df =18, $p = .208$).

Integrated Writing Task

For the integrated writing task, except for the category of "unrelated comments," the two groups of participants employed the same types of writing behaviors. Table 5.12 reviews the number and percentage of each category of the writing behaviors in the integrated writing task by the academic status of the 20 participants.

Table 5.12 *Writing Behaviors in the Integrated Writing by the Participants' Academic Status*

	Undergraduate		Graduate	
Writing behaviors	Number	%	Number	%
Commenting on one's understanding of source texts	8	0.73	13	0.99
Commenting on one's writing process	21	1.92	15	1.15
Commenting on one's writing product	101	9.23	97	7.41
Commenting on relationship between source texts	5	0.46	6	0.46
Global planning	21	1.92	27	2.06
Planning and rehearsal	196	17.92	214	16.35
Reading the instruction	15	1.37	19	1.45
Reading one's writing	118	10.79	130	9.93
Referring to notes	35	3.20	51	3.90
Referring to source texts	32	2.93	55	4.20
Revising and editing	179	16.36	261	19.94
Summarizing source texts	48	4.39	46	3.51
Analyzing the task	6	0.55	6	0.46
Verbalizing one's writing	308	28.15	369	28.19
Unrelated comments	1	0.09	0	0
Total	1094		1309	

As the table shows, in the integrated writing task, the graduate participants produced slightly more writing behaviors than their undergraduate counterparts (1309 vs. 1094), but independent *t*-test results showed that this difference was not significant. Comparing the writing behaviors across the two groups of participants, the Mann-Whitney test reported no significant differences in any of the categories. Therefore, the results demonstrated that the graduate and the undergraduate participants adopted similar writing behaviors both in terms of the types and in terms of the frequency.

Independent Writing Task

As for the comparison of the graduate and undergraduate participants for the independent writing task, the two groups once again did not differ significantly from each other in the number and type of writing behaviors produced. Table 5.13 presents the number and percentage of each category of the writing behaviors in the independent writing task by the academic status of the participants.

Table 5.13 *Writing Behaviors in the Independent Writing by the Participants' Academic Status*

	Undergraduate		Graduate	
Writing behaviors	Number	%	Number	%
Commenting on one's writing process	13	0.83	16	1.05
Commenting on one's writing product	111	7.07	86	5.65
Global planning	77	4.91	69	4.54
Planning and rehearsal	300	19.12	222	14.60
Positioning self	13	0.83	8	0.53
Reading the instruction	46	2.93	48	3.16
Reading one's writing	213	13.58	217	14.27
Referring to notes	6	0.38	12	0.79
Revising and editing	313	19.95	409	26.89
Analyzing the task	18	1.15	6	0.39
Verbalizing one's writing	458	29.20	425	27.94
Unrelated comments	1	0.06	3	0.20
Total	1569		1521	

After controlling for alpha inflation, Mann-Whitney test results revealed that the two groups of participants did not differ significantly from each other in terms of the frequency of all the writing behaviors used.

Discussion for Research Question 6

As the results illustrated, for the integrated writing task, the writing behaviors adopted by the participants did not differ significantly according to their academic experience. No significant differences in the writing behaviors failed to support the hypothesis premised on the arguments regarding the enhanced authenticity of integrated writing task (Cumming et al., 2000; Cumming et al., 2005, 2006) and findings made in previous writing process studies (Plakans, 2008; Taylor & Beach, 1984).

One possible reason is that the participants involved in this study were all matriculated ESL students in their academic programs and thus can all be regarded as advanced ESL writers. Although the time the undergraduate participants spent studying in the United States was limited (less than one year), that time period might be sufficient for them to make meaningful progress

in mastering or imitating the writing behaviors necessary to respond to the writing tasks integral to academic context at the tertiary level.

However, although, for the integrated task, the undergraduate participants' writing behaviors did not display a significant different pattern, their writing quality perceived was significantly lower than that of the graduate participants. This finding is inconsistent with the finding made in the textual analysis section of the study where no significant score difference was found. The difference might be related to the fact that the academic status was opertationalized differently in the two sections. In the textual analysis section where the ETS data was analyzed, the academic experience was determined by the self-reported data on what programs the test takers were applying to. No actual data about their academic experience was available as with the participants in the process component of the study. Given that the operationlization of the academic experience variable was more precise in the qualitative analysis, the finding that the graduate participants outperformed the undergraduate participants probably can be taken as more meaningful. It, therefore, points to the prospect that due to the more exposure and practice, the graduate participants still gained advantages with the integrated writing task in their test performance even though in terms of the writing behaviors, no significant differences were identified.

Similar to the integrated writing task, the writing behaviors elicited by the independent writing task did not differ significantly between the graduate participants and the undergraduate participants in general. Therefore, the academic experience of the participants did not seem to be related to the writing behaviors they adopted.

Summary of Qualitative Process Analysis

With regards to research question 4, through investigating the writing behaviors based on the TAP data, the study produced evidence that help to unveil what underlies the writing products in the integrated and the independent writing tasks. This information, together with the textual features revealed in research question 1, provides descriptive data to define the inherent construct of writing that is elicited by the two tasks. The results emphasized that both tasks generated writing behaviors that were recursive, a phenomenon also found with writing activities in non-testing situations. Furthermore, the differences clearly show that the writing behaviors varied with the task types according to theoretical expectations. The integrated writing task required the participants to purposefully interact with the source texts throughout the composing process while in the independent writing, the participants focused more on monitoring their understanding of the assigned topic. Similar to the findings reported in the textual analysis section, the present findings also support the inclusion of the integrated writing task as it provides an additional measure of academic English writing ability.

As for the relationship between the writing behaviors and the essay scores, the study found that none of the writing behaviors of the 20 participants varied with the essay scores in both the integrated and the independent writing. In terms of the writing behavior differences between the graduate and the undergraduate participants, the two tasks actually demonstrated a similar pattern. That is, the two groups of participants did not use significantly different writing behaviors in responding to the writing tasks. The findings, therefore, indicated that the writing behaviors did not vary with the academic experience of the participants under investigation in both the integrated and the independent writing task.

In conclusion, the qualitative process analysis indicated that the writing behaviors reported by the 20 participants did not vary with the essay scores and the academic experience of the participants but with the task type. Again, how these findings contribute to the validity argument of the two tasks will be discussed together with the findings reported in the textual analysis component in the final chapter of the dissertation.

CHAPTER 6

DISCUSSION AND CONCLUSION

In this chapter, the major findings from the quantitative textual analysis and qualitative process analysis of the TOEFL iBT integrated and independent writing will be summarized. The contribution of the findings yielded in the study to the validity of the text-based integrated writing task for the TOEFL iBT is then discussed. Implications of the findings and limitations of the current study will also be addressed. This chapter will conclude with proposed areas for future research.

Summary of the Major Findings

The study aimed to examine whether the test performance (both written products and the writing processes) vary with task type, essay scores, and academic experience of test takers in the TOEFL iBT writing section. This study yielded empirical evidence that the test performance varied with the task type in accordance with theoretical expectations. As for how the test performance related to the essay scores, the study produced mixed findings. In the textual analysis, the study demonstrated that the test takers' linguistic performance varied across score levels for both the integrated and the independent essays. Although the findings confirmed that many of proficiency descriptors listed in the rubrics can successfully predict the essays scores, some of the predictive features retained in the regression model were not captured on the scoring rubrics. Comparing the significant predictors across the integrated and the independent tasks, both similarities and differences were identified. In terms of the similarities, the study found that text length and lexical sophistication could significantly predict essay scores for both tasks. As for the differences, cohesive devices seemed to play a more important role in predicting essay scores for the integrated task than for the independent task. In analyzing the writing behaviors, however, no significant differences were found between different score groups in either the

integrated or the independent writing. Finally, in terms of the relationship between the academic experience of the test takers and their test performance (linguistic performance and writing behaviors), the study found that none of the linguistic features or the writing behaviors investigated demonstrated a significant difference between the test takers with more academic experience and those with less.

The following sections review the findings of the quantitative textual analysis followed by those of the qualitative process analysis.

Quantitative Textual Analysis

The quantitative textual analysis component of the current study examined whether the linguistic features of TOEFL iBT integrated and independent essays varied with task type, essay scores, and academic experience of test takers. A corpus of 480 TOEFL iBT essays collected in 2007 was investigated. Using Coh-Metrix, a computational textual analysis tool, the study explored linguistic features (including lexical sophistication, syntactic complexity, cohesion as well as basic text information) of the essays in relation to these three variables.

For the first research question (linguistic features in relation to task type), DA results confirmed that the linguistic features of the essays varied with the task type. The two types of essays were associated with different patterns of linguistic features, and these features were powerful enough to predict the essay type with 100% accuracy. More specifically, the integrated essays, compared with the independent essays, were found to bear more characteristics of general academic writing (detachment and structural compression as evidenced by less frequent use of personal pronoun possessive cases and more modifiers per noun phrase) and to contain more words that are concrete. On the other hand, the frequent use of personal pronoun possessive cases and logic operators suggest that the independent essays tended to be interactional and

focused on logical reasoning in arguing for the stance taken by the test takers. These differences across the two task types were also reported with the higher rated essays of the same data set and with the data set collected in 2006.

The second research questions focused on the linguistic features in relation to the essay scores within each task type. The results confirmed that linguistic performance varied with the essay scores in both the integrated and the independent essays. Regression analysis results indicated that for both types of essays, certain linguistic features can significantly predict the scores. Essay length and lexical sophistication features were found to be significant predictors of the essay scores for both the integrated and the independent essays. In accordance with the proficiency descriptors listed in the scoring rubric, semantic similarity, one of the indicators of textual cohesion, was found to be a significant predictor of the integrated essay scores while syntactic complexity was not. However, there were some discrepancies between the criteria mentioned in the scoring rubrics and the features that predicted the essay scores. First of all, for the integrated writing task, as previously mentioned, lexical sophistication features were found to have a significant effect on the essay scores even though the rubric does not list vocabulary choice as one of the evaluative criteria. For the independent writing task, although the rubric specifies that syntactic features are one of aspects to attend to in scoring, none of the features related to syntactic complexity was included in the regression model. Therefore, even though the overall results of the regression analysis confirmed that some linguistic features varied with the essay scores as theoretically expected and as the scoring rubrics stated, some differences were also noticed between the predictive linguistic features and the proficiency descriptors listed in the scoring rubrics.

Unlike the first two research questions, results from the third research question revealed that the linguistic features did not vary along with the academic experience of the test takers. This finding contradicted the expectation that more academic exposure and practice leads to different, if not better, performance. However, as mentioned previously, this finding needs to be taken with caution due to the self-reported data and the small number of the participants that could be identified as having different levels of educational experience.

To sum up, through quantitative textual analysis, the current study provided empirical evidence that the linguistic features of the TOEFL iBT essays varied with task type and score level. However, when it comes to the academic experience of the test takers, the study failed to locate significant variations in linguistic features along with that variable.

Qualitative Process Analysis

In the qualitative process analysis, the study aimed to find out whether the writing behaviors varied along with task type, essay scores, and academic experience of the participants. To answer these questions, the study examined the TAP data produced by the 20 participants involved in the think-aloud writing sessions.

Research question 4 focused on the writing behaviors in relation to task type. The study found that although the two types of writing shared many similarities, there were still differences in terms of the type and the frequency of the writing behaviors used. Both the integrated and the independent writing contained some unique writing behaviors. In the integrated writing task, the participants generated writing behaviors that were intertwined with the source texts. They were engaged with the source texts in various ways (including summarizing the source texts, commenting on the relationships between the two source texts, commenting on their understanding of the source text, etc). All these writing behaviors suggest that the participants

were not just superficially interacting with the source texts or just borrowing the content and language directly from the source texts. This finding confirmed that the use of integrated writing task encouraged the interdependent relationship between reading and writing which is prominent in academic activities at the tertiary level. Furthermore, even with the shared categories of writing behaviors, the two types of writing still demonstrated some differences. The independent writing elicited significantly more writing behaviors related to making overarching plans, and comprehending the task requirement in terms of content than the integrated writing.

As for research question 5 (writing behaviors in relation to essay scores), the study did not find that the behaviors varied along with the essay scores either for the integrated or for the independent writing. Similarly, for the last research question (writing behaviors in relation to academic status of the participants), the analysis did not reveal significant differences either.

In summary, the qualitative analysis of the TAP data revealed the writing behaviors varied with the task type. However, in terms of the relationship between the writing behaviors with the essay scores and the academic experience of the participants, the study found that the writing behaviors did not change along with these two factors.

Validity Argument for the TOEFL iBT Integrated Writing

In summary, the evidence gathered from the current study regarding the link between the expected scores and the underlying writing abilities in the TOEFL iBT writing section is mixed. In terms of the task type difference, first of all, the evidence supports the argument that the integrated writing task elicited different test performance from the independent writing task for both linguistic features and writing behaviors. The results corroborated with findings reported in previous studies that have explored writing products (Cumming et al., 2005, 2006) and writing processes (Esmaeili, 2002; Plakans, 2008, 2010; Yang, 2009) respectively. The integrated

writing task elicited different test performance from the independent writing task, indicating that the integrated task provides a different measure of academic writing ability (Huff et al., 2008). The current study, therefore, shows that the combined use of the two tasks broadens the representation of the underlying academic writing ability and thus provides justification for the addition of the integrated writing task in the writing test (Cumming et al., 2005, 2006; Huff et al., 2008). Furthermore, the descriptive information about the specific linguistic performance and writing behaviors associated with the integrated and the independent writing tasks also help to shed light on the construct inherent in each of the tasks. Both textual and writing process analyses results indicate that the integrated writing, as compared to the independent writing task, requires test takers to write in ways that more authentically resemble the types of performance needed for academic studies at the tertiary level. The integrated essays demonstrated more features of general academic writing as evidenced by less frequent use of personal pronoun possessive cases and more frequent use of linguistic features such as modifiers in noun phrases and passive voices. In terms of writing processes, the integrated writing task required the test takers to be engaged with the source texts in activities of writerly reading/listening (Church & Bereiter, 1984; Greene, 1992) or discourse synthesis (Plakans,2008, 2009) rather than superficial meaning decoding and verbatim source use. Therefore, the evidence gathered in this study also verifies and strengthens the enhanced authenticity argument of the integrated writing task (Chapelle et al., 2008).

The different skills being elicited in the two tasks also provides justification for the current practice of ETS in reporting separate scores for the integrated and the independent writing tasks rather than giving a composite score. Separate scores give more information to test takers and test users that would help them better interpret the test performance.

In terms of how the test performance related to the essay scores, the study yielded mixed results in its textual and writing process analyses. The textual analysis helped to establish the score meaning because it confirmed that the essay scores differentiated linguistic performance in both tasks, a finding often made in previous studies either on the integrated writing (Cumming et al., 2005, 2006; Gebril & Plakans, 2009) or on the independent writing (Crossley & McNamara, in press a; Frase et al., 1999; Grant & Ginther, 2000; Reppen, 1994). Furthermore, the predictive linguistic features in the regression models overlapped with many of the proficiency descriptors detailed in the scoring rubrics for both the integrated writing task and the independent writing task, thus validating that scoring rubrics. Meanwhile, it was also noticed that there was not a one-on-one correspondence between the predictive features and the proficiency descriptors detailed in the scoring rubrics. For instance, certain predictive linguistic features (lexical sophistication features in the integrated essays) were not captured by the scoring rubric, suggesting that they might co-occur with the descriptors listed in the scoring rubric or raters might attend to features not specified in the scoring guidelines. On the other hand, contrary to Yang (2009), the essay scores did not seem to differentiate writing behaviors for both the integrated and the independent writing tasks. This finding suggests that the scores do not directly reflect the use of the writing abilities (if writing behaviors are considered to be part of the writing abilities). However, the narrow range of the participants' proficiency and the relatively small number of the participants might limit the ability to discern the differences among the proficiency groups

Finally, the study did not find that academic experience of the test takers had a significant impact on the test performance in the integrated and the independent task. Test takers with more academic experience were expected to outperform those with less such experience, a hypothesis premised on the argument that writers with more practice and exposure to the academic language

and writing activities tend to be more familiar with the two tasks. The study did not support this (with the exception for the scores assigned on the integrated essays collected in the qualitative analysis). However, the findings need to be taken with caution because the operationalization of academic experience might not be very reliable in the textual analysis section. The test takers were divided into groups with different academic experience based on their self-reported data. No information about their real academic experience was available. In addition, the number of participants in the process analysis was limited and might not be large enough to show such an influence. However, the significant score difference between the participants with more experience and those with less in the process analysis definitely suggests that this deserves further attention in research.

Comparing the current study with previous studies on integrated writing, many differences can be noticed. First of all, although previous studies have also identified linguistic differences between the integrated and the independent essays (Cumming et al., 2005, 2006) and between the essays at different score levers (Cumming et al., 2005, 2006; Gebril & Plakans, 2009), it should be noted that the use of Coh-Metrix allowed a broader range of linguistic features to be investigated in the current study. Because of this reason, a more comprehensive picture was constructed as to the linguistic differences across the task types and the score levels. For instance, linguistic features such as cohesive devices and POS tags that were not explored in the previous related studies were actually found to be able to set apart the two types of essays, thus contributing to a better understanding of the differences that existed. Furthermore, even though previous studies have also found that general linguistic category of lexical sophistication varied across the task types (Cumming et al., 2005, 2006) and across the score levels (Cumming et al., 2005, 2006; Gebril & Plakans, 2009), the current study differed from the previous studies

as to what lexical sophistication features differentiated. For example, lexical diversity (measured by TTR) was reported to the one significant feature that differed across the integrated and the independent writing in Cumming et al (2005, 2006). With many more indices related to lexical sophistication being explored (including word hypernymy values, word frequency, nominalizations, etc) and statistically more rigorous measures (MDLT and D) being used to assess lexical diversity, such a finding was not made in the current study. Thirdly, unlike previous studies that either took a product or a process approach (e.g., Cumming et al., 2005, 2006; Plakans, 2008, 2010), the current study examined both the writing products and the writing processes at the same time, thus building a more comprehensive picture as to how the two tasks compared to each other.

Implications

The implications of the current study are discussed with regards to L2 writing assessment, L2 writing instruction, and use of Coh-Metrix as a textual analysis tool in L2 writing assessment respectively.

L2 Writing Assessment

This study sought to clarify whether the writing performance in the TOEFL iBT writing section varies with task type, essay scores, and academic experience of test takers in accordance with theoretical expectations. Writing performance includes not only written products but also writing processes (Cumming et al., 2006). For this reason, in this study I examined both the textual features of the essays and the writing behaviors used to complete the writing tasks. Such investigation not only contributes to a better understanding of the nature of integrated writing but also helps to clarify the link between the expected scores and the underlying writing abilities being evaluated in the TOEFL iBT writing section.

Integrated writing tasks have been promoted as an item type for their enhanced authenticity and validity. However, much of the discussion of integrated tasks is speculative and theory driven rather than empirical. Additionally, the majority of the limited studies on integrated writing, with or without comparison with independent writing, have concentrated principally on thematically-related integrated writing while little is known about text-based integrated writing. Therefore, although far from building a complete picture of integrated writing as compared to independent writing, this study does help to amass information in several important areas regarding the use of text-based integrated writing, especially when used in combination with independent writing in a test.

First of all, this study produced rich empirical data to shed more light on the inherent construct assessed by the integrated and the independent writing task, and comparison was also made across the two tasks. Through such analysis and comparison, the study provided empirical evidence showing that the two tasks did elicit different writing performance and thus affirmed the proposed rationale for the combined use of the two tasks that they help to broaden "representation of the domain of academic writing on the test" (Huff et al., 2008, p.212). The study demonstrated that compared with the independent writing task, the integrated writing task elicited meaningful interactions with the source materials on the part of the test takers, thus substantiating the strengthened authenticity argument (Cumming et al., 2005, 2006; Huff et al., 2008). The evidence is available not only from the linguistic performance perspective but also from the cognitive performance perspective. In addition, the finding that the two task types elicited different test performance also provided empirical evidence to support the argument that adding the integrated writing task diversifies the measurement of writing ability (Cumming et al., 2005; White, 1994). Therefore, the empirical evidence yielded in the current study helps to

answer the practical question of whether and why we need to use two test items simultaneously in assessing academic writing ability.

The study also investigated whether test performance varied with the writing quality perceived. The results yielded not only helped to further clarify the score meaning by illustrating the link between the observed scores and the underlying writing abilities in each of the writing tasks (Chapelle et al., 2008) but also to validate the scoring rubrics used. The current study provided empirical data showing that certain linguistic features, but not writing processes, were associated with essay scores. Furthermore, the study also explored how the test performance related to the academic experience of the test takers. Although the study failed to establish the relationship between academic experience with test performance in the integrated writing task in accordance with theoretical expectations (Chapelle et al., 2008), the integrated essay score difference identified in the writing process section does suggest that more experience might give test takers advantages in integrated writing and thus calls for further investigation. Taken together, all this information helps to clarify the link between the observed scores and the underlying writing ability being assessed (Chapelle et al., 2008), thus building a more comprehensive picture of L2 writing assessment, especially in regards to integrated writing tasks.

L2 Writing Instruction

The findings yielded in this study, especially the differences found across the integrated and the independent writing tasks, suggest that the two types of writing represent at least two different aspects of academic writing ability. Instruction in the more conventional independent argumentative writing by itself might not suffice as it does not fully prepare L2 writers for test items like text-based integrated writing tasks or more generally for the academic writing assignments that require them to compose in response to source texts.

The linguistic and cognitive differences identified in the test performance across the two tasks indicate that writing instruction and learning should include source texts and synthesis of these texts into writing to allow students with adequate exposure to such writing activities and to develop the corresponding writing ability that is integral to academic activities of higher education. Hirvela (2004) and Spack (1997) actually made a similar suggestion after viewing their students struggle with academic writing due to the lack of ability to interact with the source text(s). In fact, in addition to learning the form, Cohen (1998) also pointed out that writers can be taught to learn the writing behaviors. If that is the case, writing instruction and assignments should definitely include integrated writing tasks to show writers how to interact with source texts and how to identify and synthesize the important information from them into their own writing.

Use of Coh-Metrix

In the current study, Coh-Metrix demonstrated its effectiveness in analyzing L2 writing. With this computational textual analysis tool, L2 writing was analyzed for task difference, score difference, and test takers' academic experience difference. As mentioned previously, since textual analysis at a deeper level (such as textual cohesion) was made available through this tool, the current study was able to provide a more comprehensive picture of the nature of the linguistic reality in the TOEFL iBT integrated and independent writing tasks.

In addition, the successful application of Coh-Metrix in the current study also suggests that this tool has the potential to be another useful instrument in automated scoring of L2 writing. More specifically, the results suggest a combined use of Coh-Metrix and tools like *e-rater* in automated scoring of L2 writing since the two can provide complementary information. On one hand, through reporting features such as ill-formed verbs, pronoun errors, fragments, run-ons,

and subject-verb agreement, *e-rater* can address grammatical accuracy more directly (Attali & Burstein, 2005; Quilan, Higgins, & Wolff, 2009; Weigle, 2010), a limitation of Coh-Metrix. On the other hand, Coh-Metrix offers a range of linguistic features that are not provided by *e-rater*. These features include linguistic devices that contribute to textual cohesion (e.g., stem overlap, logical connectives, semantic similarity (LSA features), and tense and/or aspect repetition), lexical sophistication indices such as word hypernymy and polysemy values, and syntactic complexity indices including number of modifiers per noun phrase. Focusing on different areas of textual analyses, *e-rater* and Coh-Metrix, when utilized together, can enable more accurate and comprehensive evaluation of textual features in L2 writing.

Limitations

Several limitations exist for this study. First of all, this study is limited to the two writing tasks under investigation. Although in the quantitative textual analysis for the task differences, the other data set (collected in 2006) was also examined in the supplementary analysis, the majority of the study, especially the qualitative process analysis, is limited to one integrated prompt and one independent prompt. The results of the study, therefore, should be interpreted with caution that test takers might demonstrate a varied use or different types of writing performance when responding to different prompts/source texts (Yang, 2009).

Secondly, as compared to the quantitative component, the qualitative analysis was based on a simulated test, rather than a real testing condition. Therefore, whether the writing behaviors reported by the participants truly reflected what they would do in real test situations is debatable. However, it is also worth noticing that all the participants did report that they treated the writing tasks as real tests.

A further limitation related to the participants in the process analysis is that they were all matriculated ESL writers who had already spent several months studying in an English medium higher education institution. They had all met the admission requirement of the university to be matriculated. Therefore, it is questionable as to whether the data generated by them can be extended to lower proficiency writers, especially considering that in the pilot study the low proficiency writers were found to avoid the listening passage due to lack of comprehension. In addition, the small number of participants involved in the process analysis section of the study also suggests that the generalizability of the findings should be taken with caution.

Another limitation relates to the dependence on the think-aloud method to collect writing behavior data in the qualitative analysis component of the study. Although the method has been praised for being immediate and for recoding cognitive operations in real time (Swarts et al., 1984), it is a method with some recognized limitations such as the concern about the completeness of the mental activities reported (Sasaki, 2000), the distraction of the verbalization for writers (Cooper & Holzman, 1983), and the dependence on writers' verbosity (Sasaki, 2000).

Areas for Future Research

This study uncovered important information about the products and processes involved in text-based integrated and independent writing. However, in order to build a more comprehensive picture of text-based integrated writing tasks, especially with regards to the link between expected scores and the underlying ability, more work is still needed.

First of all, because the text-based integrated writing task is a newly introduced item type, it is not clear as to how test takers interpret such a task and whether and how the task interpretation relates to test performance and essay scoring. For example, the study found that in interpreting the integrated writing tasks, the participants had different opinions as to the task

expectations in terms of the format of the response. Due to the limited number of participants, no conclusion could be drawn in this study as to whether task interpretation has an impact on their test performance. Meanwhile, information related to task interpretation is an important factor to be taken in consideration for validity argument of the writing tests because the interpretation has been found to be related to writers' test performance (Ruiz-Funes, 2001).

Another important area that needs to be further and more directly addressed is whether verbatim source use has a significant influence on the test performance, especially with low proficiency writers. It is true that the text-based integrated task under investigation effectively prevents verbatim source use as it was focused on the listening passage and how it challenges the views presented in the reading passage (Enright et al., 2008). It is still of interest to find how the reading and the listening passage inform the writing of test takers in terms of the language and the format. Although in the think-aloud sessions, verbatim source use was not found to be a significant issue, it is important to be aware that this might be related to the presence of the researcher and the fact that the participants involved were all comparatively advanced writers.

Finally, grammatical accuracy of the writing products also deserves more attention. As revealed in previous studies on integrated writing, grammatical accuracy often exerts an important influence on the score assigned (Cumming et al., 2005, 2006). Due to the limitations of Coh-Metrix, the computational analysis tool utilized in the study, grammatical accuracy was not directly examined in the textual analysis section. Given that grammatical errors tend to be one of the characteristics of L2 writing (Frase et al., 1999), such information is certainly vital to a better understanding and a better use of the writing scores.

Final Remarks

This study demonstrates that text-based integrated writing tasks are a useful assessment instrument to be included in writing tests both for diversifying measurement and to promote positive washback in writing classrooms. However, so far, only limited evidence has been made available about integrated writing tasks (especially text-based integrated writing tasks), as compared to the bulk of information accumulated about independent writing tasks. Therefore, more evidence pertaining to the language use, task interpretation, cognitive operations, etc needs to be collected to provide further descriptive information about the integrated tasks and to validate such tasks in writing assessment. This is especially true given that validation of a test or a test item is an ongoing process.

REFERENCES

Ascencion, Y. (2005). *Validation of reading-to-write assessment tasks performed by second language learners.* Unpublished doctoral dissertation. Northern Arizona University.

Attali, Y. & Burstein, J. (2005). *Automated essay scoring with e-rater v. 2.0.* (TOEFL Research Rep RR-04-45). Princeton, NJ: Educational Testing Service.

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

Bachman, L. (2004). *Statistical analysis for language assessment.* Oxford, UK: Oxford University Press.

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Bereiter, C. & Scardamalia, M. (1985). Cognitive coping strategies and the problem of "inert knowledge." In S. F. Chipman, J. W. Segal, & R. Glaser (Eds.), *Thinking and learning skills: Research and open questions* (pp. 65-80). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bereiter, C. & Scardamalia, M. (1987). *The psychology of written composition.* Hillsdale, NJ: Erlbaum.

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

Biber, D.(1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Biber, D. & Gray, B. (2010). Challenging stereotypes about academic writing: complexity, elaboration, explicitness. *Journal of English for Academic Purposes, 9,* 2-20.

Brace, N., Kemp, R., & Snelgar, R. (2006). *SPSS for psychologists: a guide to data analysis using SPSS for windows* (3rd edition). London: Palgrave.

Braine, G. (1995). Writing in the natural sciences and engineering. In D. Belcher & G. Braine (Eds.), *Academic writing in a second language: essays on research and pedagogy* (pp. 113-134). Norwood, NJ: Ablex Publishing Corporation.

Brown, A. L. & Day, J. D. (1983). Macrorules for summarizing texts: the development of expertise. *Journal of Verbal Learning and Verbal Behavior, 22*, 1-14.

Brown, J. D., Hilgers, T., & Marsella, J. (1991). Essay prompts and topics: minimizing the effect of mean differences. *Written Communications, 8,* 533-556.

Camp, R. (1993). Changing the model for the direct writing assessment. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: theoretical and empirical foundations* (pp. 45-78). Cresskill, NJ: Hampton Press, Inc.

Campbell, C. (1990). Writing with other's words: using background reading text in academic compositions. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 211-230). Cambridge: Cambridge University Press.

Carson, J. (2001). A task analysis of reading and writing in academic contexts. In D. Belcher & A. Hirvela (Eds.), *Linking literacies: Perspectives on L2 reading-writing connections* (pp. 48-83). Ann Arbor, MI: The University of Michigan Press.

Carlson, S., Bridgeman, B., Camp, R., & Waanders, J. (1985). *Relationship of Admission Test Scores to Writing Performance of Native and Non-native Speakers of English*. (TOEFL Research Rep No. 19). Princeton, NJ: Educational Testing Service.

Casanave, C. & Hubbard, P. (1992). The writing assignments and writing problems of doctoral students: faculty perceptions, pedagogical issues, and needed research. *English for Specific Purposes Journal, 11*, 33-49.

Chafe, W. L. (1975). Giveness, constrastiveness, definiteness, subjects, topics, and point of view. In C.N. Li (Ed.), *Subject and topic* (pp. 26-55). New York: Academic.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). Test score interpretation and use. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the test of English as a foreign language* (pp. 145-186). NY: Routledge.

Charge, N. & Taylor, L. B. (1997). Recent development in IELTS. *English Language Teaching Journal, 51*(4), 374-380.

Chodorow, M. & Brustein, J. (2004). *Beyond essay length: evaluating e-rater's performance on TOEFL essays.* (TOEFL Research Report No.73). Princeton, NJ: ETS.

Church, E. & Bereiter, C. (1984). Reading for style. In J. M. Jensen (Ed.), *Composing and comprehending* (pp. 85-91). Urbana, IL: ERIC Clearinghouse on Reading and Communication Skills.

Cohen, A. D. (1987). Using verbal reports in research on language learning. In C. Faerch & G. Kasper (Eds.), *Introspection in second language research* (pp. 82-95). Clevedon, UK: Multilingual Matters.

Cohen, A. D. (1998). Strategies and processes in test taking and SLA. In M. H. Long & J. C. Richards (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 90-111). Cambridge: Cambridge University Press.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology, 33,* 497-505.

Connor, U. (1990). Linguistic/rhetorical measures of international persuasive student writing. *Research in the Teaching of English, 24*, 67-87.

Connor, U. & Biber, D. (1988). *Comparing textual features in high school student writing: a crosscultural study.* Unpublished manuscript.

Connor, U. & Carrell, P. (1993). The interpretation of tasks by writers and readers in holistically rated direct assessment of writing. In J. Carson & I. Leki (Eds.), *Reading in composition classroom* (pp. 141-160). Boston: Heinle and Heinle.

Cooper, M. & Holaman, M. (1983). Talking about protocols. *Collge Composition and Communication, 34*(3), 284-293.

Crossley, S. A. & McNamara, D. S. (2009). Computationally assessing lexical differences in second language writing. *Journal of Second Language Writing, 17* (2), 119-135.

Crossley, S. A. & McNamara, D. S. (in press a). Predicting second language writing proficiency: the roles of cohesion and linguistic sophistication. *Journal of Research in Reading*.

Crossley, S. A. & McNamara, D. S. (in press b). Detecting the first language of second language writers using automated indices of cohesion, lexical sophistication, syntactic complexity, and conceptual knowledge. In S. Jarvis & S. A. Crossley (Eds.), *Approaching language transfer through text classification explorations in the detect-based approach.* Bristol, United Kingdom, Multilingual Matters.

Crossley, S. A., Salsbury, T., McCarthy, P. M., & McNamara, D. S. (2008). Using latent semantic analysis to explore second language lexical development. In D. Wilson & G. Sutcliffe (Eds.), *Proceedings of the 21st International Florida Artificial Intelligence Research Society* (pp. 136-141). Menlo Park, CA: AAAI Press.

Crossley, S. A., Louwerse, M. M. , McCarthy, P.M., & McNamara, D.S. (2007). A linguistic analysis of simplified and authentic texts. *Modern Language Journal, 91*(2), 15-30.

Cumming, A. (1997). Learning to write in a second language: two decades of research. *International Journal of English Studies, 1*(2), 1-23.

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: a working paper* (TOEFL Monograph Series, Report No. 18). Princeton, NJ: ETS.

Cumming, A., Kantor, R. Baba, K., Erdoosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in writing-only and reading-to-write prototype tasks for next generation TOEFL. *Assessing Writing, 10,* 5-43.

Cumming, A., Kantor, R. Baba, K., Erdoosy, U., Eouanzoui, K., & James, M. (2006). *Analysis of discourse features iand verification of scoring levels for independent and integrated tasks for the new TOEFL* (TOEFL Monograph No. MS-30). Princeton, NJ: ETS.

Currie, P. (1998). Staying out of trouble: apparent plagiarism and academic survival. *Journal of Second Language Writing, 7*(1), 5-43.

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

Dulay, H., Burt, M. & Krashen, S. D. (1982). *Language two*. New York: Oxford University Press.

Durst, R. K. (1987). Cognitive and linguistic demands of analytical writing. *Research in the Teaching of English, 21*, 347-376.

Educational Testing Service. (2008). Validity Evidence Supporting the Interpretation and Use of TOEFL iBT™ Scores. *TOEFL iBT Research Insight, 1*(4). Retrieved May 2, 2011 from http://www.ets.org/s/toefl/pdf/toefl_ibt_insight_s1v4.pdf.

Ellis, R. (1994). *The study of second language acquisition*. Oxford: Oxford University Press.

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

Enright, M., Bridgeman, B., & Cline, F. (2002, April). Prototyping a test design for a new TOEFL. In D. Eignor (chair), *Research in support of the development of new TOEFL. Symposium conducted at the meeting of the National Council on Measurement in Education.* New Orleans. Retrieved June 20, 2010 from http://ets.org/research/conferences/aera2002.html.

Enright, M., Bridgeman, B., Eignor, D., Lee, Y. W., & Powers, D. E. (2008). Prototyping measures in listening, reading, speaking and writing. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the test of English as a foreign language* (pp. 145-186). NY: Routledge.

Ericsson, K.A. & Simon, H. (1987). Verbal reports on thinking. In C. Faerch & G. Kasper (Eds.), *Introspection in second language research* (pp. 24-53). Clevedon, UK: Multilingual Matters.

Ericsson, K.A. & Simon, H. (1993). *Protocol Analysis: verbal reports as data*. Cambridge, MA: MIT Press.

Esmaeili, H. (2002). Integrated reading and writing tasks and ESL students' reading and writing performance in an English language test. *The Canadian Modern Language Review, 58*(4), 599-622.

Faerch, C. & Kasper, G. (1987). From product to process: introspective methods in second language research. In C. Faerch & G. Kasper (Eds.), *Introspection in second language research* (pp. 5-23). Clevedon, UK: Multilingual Matters.

Feak, C. & Dobson, B. (1996). Building on the impromptu: a source-based academic writing assessment. *College ESL, 6*(1), 73-84.

Fellbaum, C. (1998). *WordNet: an electronic lexical database.* Cambridge, MA: MIT Press.

Ferris, D. (1993). The design of an automatic analysis program for L2 text research: necessity and feasibility. *Journal of Second Language Writing, 2*, 119-129.

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

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

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

Frase, L., Faletti, J., Ginther, A., & Grant, L. (1999). *Computer Analysis of the TOEFL Testof Written English.* (TOEFL Research Report No. 64). Princeton, NJ: ETS.

Friend, R. (2001). Effects of strategy instruction on summary writing of college student. *Contemporary Educational Psychology, 26*, 3-24.

Gebril, A. (2006). *Writing-only and reading-to-write academic writing tasks: a study in generalizability and test method.* Unpublished doctoral dissertation. The University of Iowa.

Gebril, A. & Plakans, L. (2009). Investigating source use, discourse features, and process in integrated writing tests. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 7*, 47-84.

Gernsbacher, M. A., & Faust, M. (1991). The mechanism of suppression: A component of general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 245-262.

Grabe, W. (2001). Reading-writing relations: theoretical perspectives and instructional practices. In D. Belcher & A. Hirvela (Eds.), Linking literacies: perspectives on L2 reading-writing connections (pp. 15-47). Ann Arbor: University of Michigan Press.

Grabe, W. & Kaplan, R. B. (1996). *Theory and practice of writing: an applied linguistic perceptive.* London: Longman.

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

Graesser, A. C., Jeon, M., Yang, Y., & Cai, Z. (2007). Discourse cohesion in text and tutorial dialogue. *Information Design Journal, 15,* 199-213.

Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text? In A. P. Sweet & C. E. Snow (Eds.), *Rethinking reading comprehension* (pp. 82-98). New York: Guilford.

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

Green, A. (1998). *Verbal protocol analysis in language testing research: a handbook*. Cambridge: Cambridge University Press.

Greene, S. (1992). Mining texts in reading to write. *Journal of Advanced Composition, 12*(1), 151-170.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). *Multivariate data analysis.* New York: Macmillan.

Hale, G. A., Taylor, C., Bridgeman, B., Carson, J., Kroll, B., & Kantor, R. (1996). *A study of writing tasks assigned in academic degree programs* (TOEFL Research Report No. 54). Princeton, NJ: Educational Testing Service.

Halliday, M.A.K. (1994). Foreword. *Functions of Language*, 1(1), 1–5.

Halliday, M. A. K. & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Hamp-Lyons, L. (1991). Introduction. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic context* (pp. 1-4). Norwood, NJ: Ablex.

Hamp-Lyons, L. & Kroll, B. (1996). Issues in ESL writing assessment: an overview. *College ESL, 6*(1), 52-72.

Hayes, J. R. & Flower, L. (1980). Identifying the organization of writing processes. In L. Gregg & E. Steinberg (Eds.), *The science of writing* (pp. 1-27). Mahwah, NJ: Lawrence Erlbaum Associates.

Hinkel, E. (2002). *Second language writers' text: Linguistic and rhetorical features.* Mahwah, NJ: Lawrence Erlbaum.

Hirvela, A. (2004). *Connecting Reading and Writing in Second Language Writing Instruction.* Ann Arbor: The University of Michigan Press.

Homburg, T. J. (1984). Holistic evaluation of ESL compositions: can it be validated objectively? *TESOL Quarterly, 18*, 87-107.

Horowitz, D. (1991). ESL writing assessments: contradictions and resolutions. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic context* (pp. 71-86). Norwood, NJ: Ablex.

Huff, K., Powers, D. E., Kantor, R. N., Mollaun, P., Nissan, S., & Schedl, M. (2008). Prototyping a new test. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the test of English as a foreign language* (pp. 187-225). NY: Routledge.

Jamieson, J. M., Eignor, D., Grabe, W., & Kunnan, A. J. (2008). Framework for a new TOEFL. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the test of English as a foreign language* (pp. 145-186). NY: Routledge.

Jarvis, S. (in press). Data mining with learner corpora: Choosing classifiers for L1 detection. In F. Meunier, S. De Cock, G. Gilquin, M. Paquot (Eds.), *A Taste for Corpora: In honour of Sylviane Granger.* Amsterdam & Philadephia: John Benjamins.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing, 19,* 57-84.

Jennings, M., Fox, J., Graves, B., & Shohamy, E. (1999). The test taker's choice: An investigation of the effect of topic on language test performance. *Language Testing, 16*(4), 426-456.

Johns, A. & Mayes, P. (1990). An analysis of summary protocols of University ESL students. *Applied Linguistics, 11,* 253-271.

Jurafsky, D. & Martin, J. H. (2008). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition.* Upper Saddle River, NJ: Prentice Hall.

Kennedy, M. L. (1985). The composing process of college students writing from sources. *Written Communication, 2,* 434-456.

Kintsch, W. & van Dijk, T. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363-394.

Kormos, J. (1998). The use of verbal reports in L2 research: verbal reports in L2 speech production research. *TESOL Quarterly, 32,* 353-358.

Kroll, B. (1979). A survey of writing needs of foreign and American college freshmen. *English Language Teaching Journal, 33*(3), 219-226.

Lecocke, M. & Hess, K. (2006). An empirical study of unvariate and genetic algorithm-based feature selection in binary classification with microarray data. *Cancer Informatics, 2,* 313-327.

Lee, D. & Anderson, C. (2007). Validity and topic generality of a writing performance test. *Language Testing, 24*(3), 307-330.

Leki, I. & Carson, J. (1997). Completely different worlds: EAP and the writing experiences of ESL students in university courses. *TESOL Quarterly, 31,* 36-69.

Lewkowicz, J. (1994). *Writing from sources: does source material help or hinder students' performance?* Paper presented at the Annual International Language in Education Conference, Hong Kong. [ERIC Document Reproduction Service No. ED386050].

Lewkowicz, J. (1997). *Investigating authenticity in language testing.* Unpublished doctoral dissertation. University of Lancaster.

Lightman, E.J., McCarthy, P.M., Dufty, D.F., & McNamara, D.S. (2007). The structural organization of high school educational texts. In D. Wilson & G. Sutcliffe (Eds.),

Proceedings of the twentieth International Florida Artificial Intelligence Research Society Conference (pp. 235-240). Menlo Park, California: The AAAI Press.

Lumley, T. (2005). *Assessing second language writing: the raters' perspective.* Frankfurt, Germany: Peter Lang.

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

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42,* 381–392.

McCarthy, P. M., Lewis, G. A., Dufty, D. F., & McNamara, D. S. (2006). Analyzing writing styles with Coh-Metrix. In G. C. J. Suteliffe & R. G. Goebel (Eds.), *Proceedings of the 19th Annual Florida Artificial Intelligence Research Society International Conference* (FLAIRS, pp. 764-770). Melbourne Beach, FL: AAAI Press.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communications, 27*(1), 57-86.

McNamara, D. S. & Graesser, A. C. (in press). Coh-Metrix: An automated tool for theoretical and applied natural language processing. In P. M. McCarthy & C. Boonthum (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution.* Hershey, PA: IGI Global.

McNamara, D.S. & McDaniel, M. (2004). Suppressing irrelevant information: Knowledge activation or inhibition? *Journal of Experimental Psychology: Learning, Memory, & Cognition, 30,* 465–482.

McNamara, D.S., Ozuru, Y., Graesser, A.C., & Louwerse, M. (2006). Validating Coh-Metrix. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 573). Mahwah, NJ: Eribaum.

Menard, S. (1995). *Applied logistic regression analysis: Sage University series on quantitative applications in the social sciences.* Thousand Oaks, CA: Sage.

Messer, S. D. (1997). *Evaluating ESL written summaries: an investigation of the ESL integrated summary profile (ISP) as a measure of the summary writing ability of ESL students.* Unpublished doctoral dissertation. Florida State University.

Moss, P. A. (1994). Validity in high stakes writing assessment: problems and possibilities. *Assessing Writing, 1*(1), 109-128.

Murray, D.H. (1982). *Learning by teaching.* Montclair, NJ: Boynton/Cook.

Nation, P. (1988). *Word lists.* Victoria: University of Wellington Press.

Neter, J., Wasserman, W., & Kutner, M. H. (1989). *Applied linear regression models.*

Homewood, IL: Irwin.

O'Brien, R. T. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & quantity, 41,* 673-690.

Pearson, P. D. (1974-1974). The effects of grammatical complexity on children's comprehension, recall, and conception of certain semantic relationships. *Reading Research Quarterly, 10,* 155-192.

PErfetti, C.A., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension skill. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: a handbook* (pp. 227-247). Oxford: Blackwell.

Plakans, L. (2007). *Second language writing and reading-to-write assessment tasks: a process study.* Unpublished doctoral dissertation. The University of Iowa.

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

Plakans, L. (2010). Independent vs. integrated writing tasks: A comparison of task representation. *TESOL Quarterly, 44*(1), 185-194.Quilan, T., Higgins, D., & Wolff, S. (2009). *Evaluating the construct-coverage of the e-rater scoring engine* (ETS Research Report RR-09-01). Princeton, NJ: ETS.

Quirk, R., Greenbaum,S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language.* London: Longman.

Rayner, K., & Pollatsek, A. (1994). *The psychology of reading.* Mahwah, NJ: Lawrence Erlbaum Associate.

Reid, J. (1990). Responding to different topic types: a quantitative analysis from a contrastive rhetoric perspective. In B. Kroll (Ed.), *Second Language Writing: research insights for the classroom* (pp. 191-210). Cambridge: Cambridge University Press.

Reid, J. (1986). Using the Writer's Workbench in composition teaching and testing. In C. Stansfield (Ed.), *Technology and language testing* (pp. 167-188). Alexandria,VA: TESOL.

Reid. J. (1992). A computer text analysis of four cohesion devices in English discourse by native and nonnative writers. *Journal of Second Language Writing, 1*, 79-107.

Reppen, R. (1994). *Variation in elementary student language: a multi-dimensional perspective.* Unpublished doctoral dissertation. Northern Arizona University.

Ruiz-Funes, M. (2001). Task representation in foreign language reading-to-write. *Foreign Language Annals, 34*(3), 226–234.

Sasaki, M. (2000). Toward an empirical model of EFL writing processes: an exploratory study. *Journal of Second Language Writing, 9*, 259-291.

Song, M. Y. (2007).*A correlational study of the holistic measure with the index measure of accuracy and complexity in international English-as-a-second-language (ESL) student writings.* Unpublished doctoral dissertation. University of Mississippi.

Spivey, N. (1984). *Discourse synthesis: constructing texts in reading and writing.* Newwark, DE: International Reading Association.

Spivey, N. (1997). *The constructivist metaphor: Reading, writing, and making of meaning.* San Diego, CA: Academic Press.

Stratman, J. & Hamp-Lyons, L. (1994). Reactivity in concurrent think-aloud editing protocols: issues for research. In P. Smagorinsky (Ed.), *Speaking about writing: reflections on research methodology* (pp. 89-112). LA, CA: Sage Publications.

Swarts, H., Flower, L.S., & Hayes, J. R. (1984). Designing protocol studies of the writing process: an introduction. In R. Beach & L.S. Bridwell (Eds.), *New directions in composition research* (pp. 53-71). New York: Guilford Press.

Taylor, B. & Beach, R. W. (1984). The effects of text structure instruction on middle-grade students' comprehension and production of expository text. *Reading Research Quarterly, 19,* 134-146.

Toglia, M. P. & Battig, W. R. (1978). *Handbook of semantic word norms.* New York: Erlbaum.

Trites, L. & McGroarty, M. (2005). Reading to learn and reading to integrate: new tasks for reading comprehension test? *Language Testing, 22,* 174-210.

van de Kopple, D. (1985). Some exploratory discourse on metadiscourse. *College Composition and Communication, 36,* 82-95.

van Dijk, T. A. & Kintsch, W. (1977). Cognitive psychology and discourse. In W. U. Dressler (Ed.), *Current trends in text linguistics* (pp. 61-80). Berlin/ New York: de Gruyter.

Wallace, C. (1997). IELTS: global implications of curriculum and materials design. *ELT Journal, 51,* 370-373.

Watanabe, Y. (2001). *Read-to-write tasks for the assessment of second language academic writing skills: investigating text features and rater reactions.* Unpublished doctoral dissertation. University of Hawaii.

Weir, C. (1983). *Identifying the language needs of overseas students in tertiary education in the United Kingdom.* Unpublished PhD Thesis, University of London Institute of Education.

Weigle, S. C. (2002). *Assessing Writing*. New York, NY, Cambridge University Press.

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

Weigle, S.C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing*, *27*(3), 335-353.

White, E. (1994). *Teaching and assessing writing,* 2nd ed. San Francisco, CA: Jossey-Bass Publishers.

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

Yang, H. C. (2009). *Exploring the complexity of second language writers' strategy use and performance on an integrated writing test through structural equation modeling and qualitative approaches.* Unpublished doctoral dissertation. The University of Texas at Austin.

Yang, L. & Shi, L. (2003). Exploring six MBA students' summary writing by introspection. *Journal of English for Academic Purposes, 2*, 165-192.

Zipf, G. K. (1945). The meaning-frequency relationship of words. *Journal of General Psychology, 33*, 251–256.

APPENDIX A

SCORING RUBRICS

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

Score	Task Description
5	A response at this level successfully selects the important information from the lecture and coherently and accurately presents the information in relation to the relevant information presented in the reading. The response is well organized, and occasional language errors that are present do not result in inaccurate or imprecise presentation of content or connections.
4	A response at this level is generally good in selecting the important information from the lecture and in coherently and accurately presenting this information in relation to the relevant information in the reading, but it may have minor omission, inaccuracy, vagueness, or imprecision of some content from the lecture or in connection to points made in the reading. A response is also scored at this level if it has more frequent or noticeable minor language errors, as long as such usage and grammatical structures do not result in anything more than an occasional lapse of clarity or in the connections of ideas.
3	A response at this level contains some important information from the lecture and conveys some relevant connection to the reading, but it is marked by one or more of the following: • Although the overall response is definitely oriented to the task, it conveys only vague, global, unclear, or somewhat imprecise connection of the points made in the lecture to points made in the reading. • The response may omit one major key point made in the lecture. • Some key points made in the lecture or the reading, or connections between the two, may be incomplete, inaccurate, or imprecise. • Errors of usage and/or grammar may be more frequent or may result in noticeably vague expressions or obscured meanings in conveying ideas and connections.
2	A response at this level contains some relation information from the lecture, but is marked by significant language difficulties or by significant omission or inaccuracy of important ideas from the lecture or in the connections between the lecture and the reading; a response at this level is marked by one or more of the following: • The response significantly misrepresents or completely omits the overall connection between the lecture and the reading. • The response significantly omits or significantly misrepresents important points made in the lecture. • The response contains language errors or expressions that largely obscure connections or meaning at key junctures, or that would likely obscure understanding of key ideas for a reader not already familiar with the reading and the lecture.
1	A response at this level is marked by one or more of the following: • The response provides little or no meaningful or relevant coherent content from the lecture.

	• The language level of the response is so low that it is difficult to derive meaning.
0	A response at this level merely copies sentences from the reading, rejects the topic or is otherwise not connected to the topic, is written in a foreign language, consists of keystroke characters, or is blank.

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

Score	Task Description
5	An essay at this level largely accomplishes all of the following: • Effectively addresses the topic and task • Is well organized and well developed, using clearly appropriate explanations, exemplifications, and/or details • Display unity, progression, and coherence • Displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice, and idiomaticity, though it may have minor lexical or grammatical errors.
4	An essay at this level largely accomplishes all of the following: • Addresses the topic and task well, though some points may not be fully elaborated. • Is generally well organized and well developed, using appropriate and sufficient explanations, exemplifications, and/or details • Displays unity, progression, and coherence, though it may contain occasional redundancy, digression, or unclear connections • Displays facility in the use of language, demonstrating syntactic variety and range of vocabulary, though it will probably have occasional noticeable minor mistakes in structure, word form, or use of idiomatic language that do not interfere with meaning
3	An essay at this level is marked by one or more of the following: • Addresses the topic and task using somewhat developed explanations, exemplifications, and/or details • Displays unity, progression, and coherence, though connection of ideas may be occasionally obscured • May demonstrate inconsistent facility in sentence formation and word choice that may result in lack of clarity and occasionally obscure meaning • May display accurate but limited range of syntactic structures and vocabulary
2	An essay at this level reveal one or more of the following weaknesses: • Limited development in response to the topic and task • Inadequate organization or connection of ideas • Inappropriate or insufficient exemplifications, explanations, or details to support or illustrate generalizations in response to the task • A noticeably inappropriate choice of words or word forms • An accumulation of errors in sentence structure and/or usage
1	An essay at this level is seriously flawed by one or more of the following weaknesses: • Serious disorganization or underdevelopment

	• Little or no detail, or irrelevant specifics, or questionable responsiveness to the task • Serious and frequent errors in sentence structure or usage
0	An essay at this level merely copy words from the topic, rejects of the topic, or is otherwise not connected to the topic, is written in a foreign language, consists of keystroke characteristics, or is blank.

APPENDIX B

Recruitment Flyer

Participants Wanted!

If you are:

a. GSU undergraduate or graduate student above 18 years old
b. A non-native English speaker (English is not your first language)
c. Did not earn any academic degree (including high school degree) in an English speaking country

You are welcome to participate in a research study on English writing.

Purpose of the study: To compare two different writing tasks (writing based on source materials and writing based on prompt only)

Participants will be asked to:

a. write a response essay based on outside sources
b. Read a prompt and write a response essay
c. Fill out questionnaires about your background and thoughts during the writing tasks
d. Complete interviews about your writing experience.

If you are interested, please contact

Liang Guo email address telephone number

The participant will be videotaped while composing the two essays. The study will require each participant to come in for 1 visit for about 2.5 hours depending on your speed. Each participant will receive $50 in cash. Also you may request a copy of your own writing samples.

Investigators:

Liang Guo, Doctoral candidate, Department of Applied Linguistics and ESL, GSU

Professor Sara Weigle, Ph.D., Department of Applied Linguistics and ESL, GSUAPPENDIX C

APPENDIX C

IRB APPROVAL LETTER

INSTITUTIONAL REVIEW BOARD

Mail: P.O. Box 3999
Atlanta, Georgia 30302-3999
Phone: 404/413-3500
Fax: 404/413-3504

In Person: Alumni Hall
30 Courtland St, Suite 217

February 25, 2011

Principal Investigator: Weigle, Sara C

Student PI: Liang Guo

Protocol Department: Applied Linguistics & ESL

Protocol Title: Process in TOEFL iBT Independent and Integrated Writing Tasks: An Investigation of Construct Validity

Submission Type: Protocol H11311

Review Type: Expedited Review

Approval Date: February 25, 2011

Expiration Date: February 24, 2012

The Georgia State University Institutional Review Board (IRB) reviewed and approved the above referenced study and enclosed Informed Consent Document(s) in accordance with the Department of Health and Human Services. The approval period is listed above.

Federal regulations require researchers to follow specific procedures in a timely manner. For the protection of all concerned, the IRB calls your attention to the following obligations that you have as Principal Investigator of this study.

1. When the study is completed, a Study Closure Report must be submitted to the IRB.

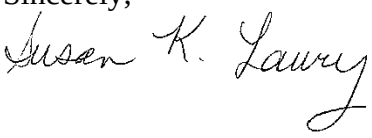
2. For any research that is conducted beyond the one-year approval period, you must submit a Renewal Application 30 days prior to the approval period expiration. As a courtesy, an email reminder is sent to the Principal Investigator approximately two months prior to the expiration of the study. However, failure to receive an email reminder does not negate your responsibility to submit a Renewal Application. In addition, failure to return the Renewal Application by its due date must result in an automatic termination of this study. Reinstatement can only be granted following resubmission of the study to the IRB.

3. Any adverse event or problem occurring as a result of participation in this study must be reported immediately to the IRB using the Adverse Event Form.

4. Principal investigators are responsible for ensuring that informed consent is obtained and that no human subject will be involved in the research prior to obtaining informed consent. Ensure that each person giving consent is provided with a copy of the Informed Consent Form (ICF). The ICF used must be the one reviewed and approved by the IRB; the approval dates of the IRB review are stamped on each page of the ICF. Copy and use the stamped ICF for the coming year. Maintain a single copy of the approved ICF in your files for this study. However, a waiver to obtain informed consent may be granted by the IRB as outlined in 45CFR46.116(d).

All of the above referenced forms are available online at https://irbwise.gsu.edu. Please do not hesitate to contact Susan Vogtner in the Office of Research Integrity (404-413-3500) if you have any questions or concerns.

Sincerely,



Susan Laury, IRB Chair

Federal Wide Assurance Number: 00000129

APPENDIX D

INFORMED CONSENT FORM0

Georgia State University
Department of Applied Linguistics
Informed Consent

Title: Process in TOEFL iBT Independent and Integrated Writing Tasks: An Investigation of Construct Validity

Student researcher: Liang Guo
Faculty supervisor: Sara Cushing Weigle

I. Purpose:

You are invited to participate in a research study. The goal of this research study is to compare the writing processes in two writing tasks (prompt-only and source-based writing tasks). You are asked to participate because you are an English as a second language student. A total of 20 participants (10 undergraduate and 10 graduate students) will be asked to participate in this study. Participation will require about 2.5 hours of your time.

II. Procedures:

If you decide to participate, you will participate in 2.5 hours of activities related to essay writing. You will first be asked to complete a questionnaire about your demographic information and essay writing experience. You will then write two essays and meanwhile verbally report every thought. The essay writing process will be videotaped. The student investigator will also interview you for your essay writing processes.

III. Risks:

In this study, you will not have any more risks (dangers) than you would in a normal day of life.

IV. Benefits:

Participation in this study may or may not benefit you personally. Overall, we hope to gain information about your writing processes in composing the prompt-only and source-based writing tasks.

V. Voluntary Participation and Withdrawal (decision to stop participating):

Participation in research is voluntary. You do not have to be in this study. If you decide to be in the study and change your mind, you have the right to drop out at any time. You can stop participating at any time. Whatever you decide, you will not lose any benefits to which you are otherwise entitled.

VI. Confidentiality:

We will keep your records private to the extent allowed by law. We will use a fake name rather than your name on study records. Only the faculty and student investigators will be able to look at the information you provide. The information may also be shared with those who make sure the study is done correctly (GSU Institutional Review Board and/or the Office for Human Research Protection). This information will be stored in a locked

computer with firewall protection. The video recordings with also be stored in digital format on the same computer. The videotapes will be stored in a locked cabinet which only the student investigator has access to. Your name and other facts that might point to you will not appear when we present this study or publish its results. The findings will be summarized and reported in group form. You will not be identified personally. The collected data may be used for future study and analysis as well. In the event that this occurs, your identity will remain completely anonymous.

VII. Contact Persons:

Contact Sara Weigle at 404- 413-5192 or Liang Guo at eslligx@langate.gsu.edu if you have questions about this study. If you have questions or concerns about your rights as a participant in this research study, you may contact Susan Vogtner in the Office of Research Integrity at 404-413-3513 or svogtner1@gsu.edu.

VIII. Copy of Consent Form to Subject:

We will give you a copy of this consent form to keep.

If you are willing to volunteer for this research and be video recorded, please sign below.

_____	_____
Participant	Date
_____	_____
Principal Investigator or Researcher Obtaining Consent	Date

APPENDIX E

BACKGROUND QUESTIONNAIRE

- Gender: Male _____ female _____
- Home country: _____
- Native language: _____
- Academic status: Graduate _____ Undergraduate _____
- Major _____

English experience

How many months have you studied in the U.S.? _____ months

Writing courses and experiences

- Have your taken English writing courses in your home country? Yes _____ No _____
- If yes, please specify what kinds of writing courses have you taken? (Choose all that apply)

a. English composition course at your undergraduate university
b. English composition course in your graduate program
c. TOEFL writing test preparation course
d. English composition course in high school
e. Others please specify _____

- Have you taken English writing courses in the U.S.? Yes _____ No _____

- If yes, please specify what kinds of writing courses have you taken? (Choose all that apply)

a. English composition course at your undergraduate university in the U.S.
b. English composition course in your graduate program in the U.S.
c. ESL writing course in an ESL program in the U.S.
d. TOEFL writing test preparation course in the U.S.
e. Others please specify _____

- What types of writing have you done in your English writing or academic courses? (choose all that apply)

a. Expository essays (e.g., compare and contrast, cause and effect essays, etc)
b. Descriptive essays (e.g., description of an object, place, experience, etc)
c. narrative essays (e.g., tell a story)
d. argumentative essays (e.g., choose a position and provide examples and details to back it up)
e. lab reports

f. summaries
g. research papers (articles including introduction, literature review, methods, results, etc)

Opinions about writing

	Totally Disagree	Partially Disagree	Neither Agree or Disagree	Partially Agree	Totally Agree
I enjoy writing in English	_____	_____	_____	_____	_____
I have strong English writing skills.	_____	_____	_____	_____	_____

TOEFL experience

- If you have taken TOEFL, which year did you take it last? _____
- Which form of TOEFL did you take? Paper-based _____ computer-based_____ internet-based_____
- What was your most recent TOEFL score_____
- TOEFL sub-scores: writing _____ grammar _____ listening _____ reading _____ speaking _____

(If you cannot remember your exact scores, please make your best guess)

APPENDIX F

TAP TRAINING SHEET

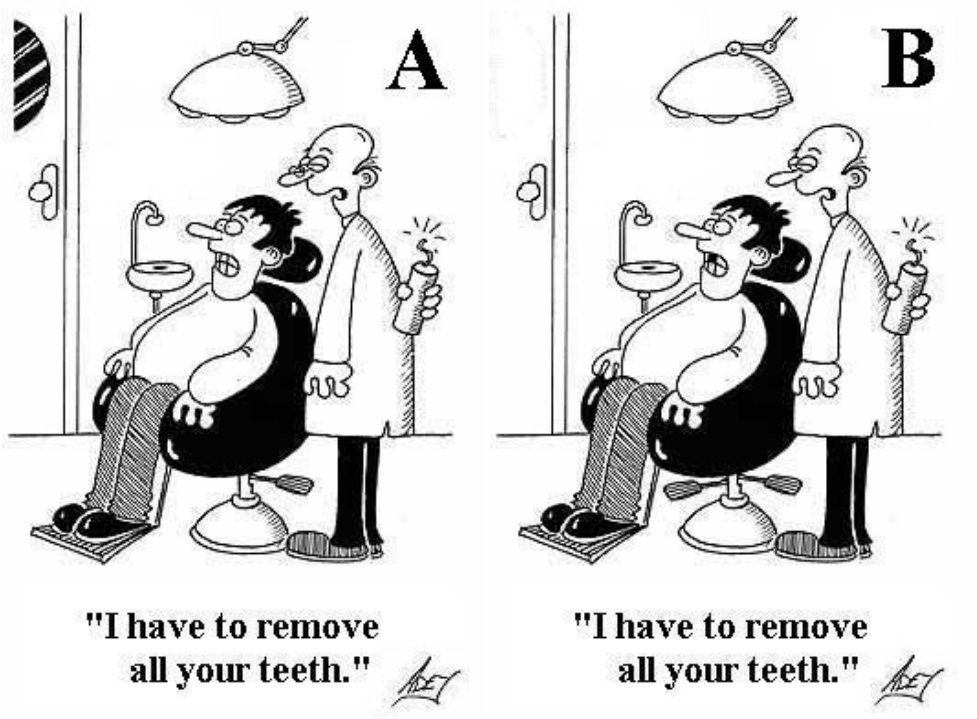
Think-aloud Protocol

Instruction:

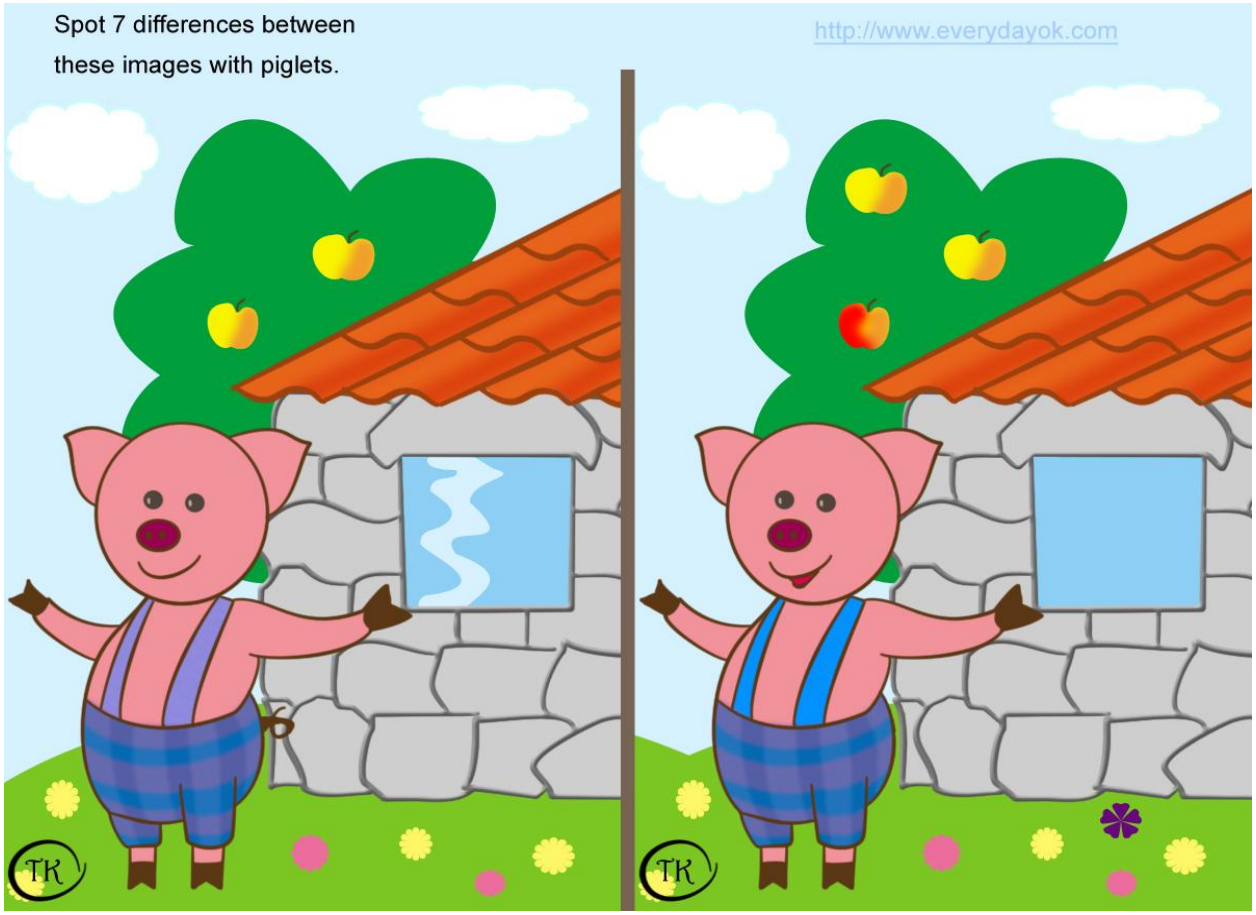
In this study, I am interested in what you think about as you perform the writing tasks that I give you. To do this, I will ask you **to think aloud** as you write. By "think aloud", I mean that I want you to say out loud **everything** that you say to yourself silently as you write. I would like you to talk continuously from the time you start reading the prompt until you finish writing. It is very important that you keep talking and articulating everything that goes through your mind. Just act as you are alone in the room speaking to yourself. It is very important that you keep talking. If you are silent for any length of time, I will remind you to keeping talking aloud. I will ask you to use English as you talk, but if you need to use your native language occasionally to avoid interruption, please do so.

Your performance will be video-recorded. I might also interview you about your writing session after you finish the task. Before we proceed with the experiment, we will start with a couple of practice items to get you familiar with think-aloud practice.

First, I will show you an example of how I would talk aloud while completing a "spot 5 differences between 2 pictures" task.



Now I would like you to practice think aloud as you solve the following "spot 7 differences between 2 pictures" task in your head.



Please go ahead and speak aloud your thinking process while solving the problem.

Now I will give you one more practice item. I want you to do the same thing for this task. I want you to speak aloud everything that goes through your mind. Any questions?

Here is your next task. It is a writing task.

Suppose that you are a student in an English writing class. You will be absent from the next class because you caught a cold. Please write an email to your professor (Dr. Crawford) to let her know your absence and ask for class assignments. You can use your own name.

Please use the computer to write the email. The email address of the professor is provided.

(Please be aware that I might say "please keep talking" if you fall silent for more than 20 seconds.)

APPENDIX G

POST-TASK QUESTIONNAIRE-INTEGRATED WRITING TASK

Please respond to each statement below using a check (√). Choose from: Strongly agree, Agree, Maybe, Disagree, or Strongly Disagree)

	Strongly agree	**Agree**	**Maybe**	**Disagree**	**Strongly Disagree**
The integrated writing task was a good test of my ability to read in English.					
The integrated writing task was a good test of my ability to write in English.					
The integrated writing task was a good test of my knowledge of English grammar.					
The integrated writing task was a good test of my ability to use English grammar correctly.					
The reading passage was interesting.					
The reading passage was easy to understand.					
The listening comprehension passage was interesting.					
The listening comprehension passage was easy to understand.					
It was easy to think of what to write in the writing part of the test/the integrated writing task.					
I treated the integrated writing task like a real test.					
The integrated writing task was easier than the independent writing task.					
Thinking-aloud affected the way I wrote the essay.					
Thinking-aloud made me aware of things I had not thought about before regarding writing integrated essays.					
It was easier to think aloud with the integrated writing task than with the independent writing task.					

APPENDIX H

POST-TASK QUESTIONNAIRE-INDEPENDENT WRITING TASK

Please respond to each statement below using a check (√). Choose from: Strongly agree, Agree, Maybe, Disagree, or Strongly Disagree)

	Strongly agree	**Agree**	**Maybe**	**Disagree**	**Strongly Disagree**
The independent writing task was a good test of my ability to write in English.					
The independent writing task was a good test of my knowledge of English grammar.					
The independent writing task was a good test of my ability to use English grammar correctly.					
It was easy to think of what to write in the writing part of the test/the independent writing task.					
I treated the independent writing task like a real test.					
The independent writing task was easier than the integrated writing task.					
Thinking-aloud affected the way I wrote the essay.					
Thinking-aloud made me aware of things I had not thought about before regarding writing independent essays.					
It was easier to think aloud with the independent writing task than with the integrated writing task.					

APPENDIX I

SEMI-STRUCTURED INTERVIEW

- Why do you think the source-based (or prompt-only) writing task was more difficult?

  In what way?

- (optional) You mentioned that you did not treat the independent writing task like a real test, what would you have done differently if the essay you wrote have been a real test?

- (optional) You mentioned that you did not treat the integrated writing task like a real test, what would you have done differently if the essay you wrote have been a real test?

- What are your thoughts about thinking aloud?

- (optional) You mentioned that thinking aloud affected the way you wrote the essays. Why/how?

- I noticed that you _____ in the _____. Can you please explain why you did that?

- Do you have any thoughts or comments that you would like to add about your experience of thinking aloud while writing the essays?

- Other questions regarding the writing processes (stimulated recall interviews)

APPENDIX J

DIRECTIONS, PROMPT, AND SOURCE TEXTS FOR THE INTEGRATED WRITING TASK

Writing Section Directions (Overview)

This section measures your ability to use writing to communicate in an academic environment. There will be two writing tasks. For the first writing task, you will read a passage, listen to a lecture, and then answer a question based on what you have read and heard. For the second writing task you will answer a question based on your own knowledge and experience.

Writing Section Directions (Question 1)

For this task, you will read a passage about an academic topic. A clock at the top of the screen will show how much time you have to read. You may take notes on the passage while you read. The passage will then be removed and you will listen to a lecture about the same topic. While you listen you may also take notes. You will be able to see the reading passage again when it is time for you to write. You may use your notes to help you answer the question. You will then have to write a response to a question that asks you about the relationship between the lecture you heard and the reading passage. Try to answer the question as completely as possible using information from the reading passage and the lecture. The question does not ask you to express your personal opinion.

Your response will be judged on the quality of your writing, and on the completeness and accuracy of the content. Immediately after the reading time ends the lecture will begin, so keep your headset on until the lecture is over.

Reading:

Since the 1960s, fish farming—the growing and harvesting of fish in enclosures near the shoreline—has become an increasingly common method of commercial fish production. In fact, almost one third of the fish consumed today are grown on these farms. Unfortunately fish farming brings with it a number of harmful consequences and should be discontinued.

One problem with fish farming is that it jeopardizes the health of wild fish in the area around the farm. When large numbers of fish are confined to a relatively small area like the enclosures used in farming, they tend to develop diseases and parasitic infections. Although farmers can use medicines to help their own fish, these illnesses can easily spread to wild fish in the surrounding waters, and can endanger the local populations of those species.

In addition, farm-raised fish may pose a health risk to human consumers. In order to produce bigger fish faster, farmers often feed their fish growth-inducing chemicals. However, the

effects of these substances on the humans who eat the fish have not been determined. It is quite possible that these people could be exposed to harmful or unnatural long-term effects.

A third negative consequence of fish farming relates to the long-term wastefulness of the process. These fish are often fed with fish meal, a food made by processing wild fish. Fish farmers must use several pounds of fish meal in order to produce one pound of farmed fish. So producing huge numbers of farm-raised fish actually reduces the protein available from the sea.

Listening:

Now, listen to part of a lecture on the topic you just read about.

Audio

(Professor) The reading passage makes it seem that fish farming is a reckless, harmful enterprise. But each of the arguments the reading passage makes against fish farming can be rebutted.

(Professor) First, what are the wild, local fish that fish farms are supposed to harm? The fact is that in many coastal areas, local populations of wild fish were already endangered – not from farming, but from traditional commercial fishing. Fish farming is an alternative to catching wild fish. And with less commercial fishing, populations of local species can rebound. The positive effect of fish farming on local, wild fish populations is much more important than the danger of infection.

(Professor) Second, let's be realistic about the chemicals used in fish farm production. Sure, farmers use some of these substances. But the same can be said for most of the poultry, beef, and pork that consumers eat. In fact, rather than comparing wild fish with farm fish as the reading does, we should be comparing the consumption of fish with the consumption of these other foods. Fish has less fat and better nutritional value than the other farm-raised products, so consumers of farm-raised fish are actually doing themselves a favor in terms of health.

(Professor) Finally, the reading makes claims that fish farming is wasteful. It's true that some species of farm-raised fish are fed fishmeal. But the species of fish used for fishmeal are not usually eaten by humans. So fish farming is a way of turning inedible fish into edible fish. Contrary to what the reading says, fish farming increases the number of edible fish, and that's what's important.

Question:

Summarize the points made in the lecture, being sure to explain how they challenge the specific points made in the reading passage.

APPENDIX K

DIRECTIONS AND PROMPT FOR THE INDEPENDENT WRITING TASK

Writing Section Directions

In this section you will demonstrate your ability to write an essay in response to a question that asks you to express and support your opinion about a topic or issue. The question will be presented on the next screen and will remain available to you as you write.

Your essay will be scored on the quality of your writing. This includes the development of your ideas, the organization of your essay, and the quality and accuracy of the language you use to express your ideas. Typically an effective essay will contain a minimum of 300 words. You will have **30 minutes** to plan, write, and revise your essay. If you finish your response before time is up, you may click on **Next** to end this section.

Question:

Do you agree or disagree with the following statement? In today's world, the ability to cooperate well with others is far more important than it was in the past. Use specific reasons and examples to support your answer.

APPENDIX L

SAMPLE ESSAYS WITH COH-MEXTRIX INDEX SCORES

Integrated Essay (20073264; 5 points)

The reading passage gave an impression that fish farming was harmful and useless. However, the points mentioned are insufficient and can be beaten by the actual facts each and every.

First, the passage mentioned that fish farming was harmful to wild fish in the same area. The fact is that fish farming actually save the lifes of wild fish which has already been endangered by overfishing. Fish farming provides an alternative fish supply and eventually gives the opportunity for wild fish population to grow.

Secondly, fish farming gives people a chance to eat healthy food instead of harming their health. Not to mention the wild fish consumption, people eat raised chicken and beef with the chemically mentioned anyways. Fish has a lower fat rate and contains more nutrients than any other meat above. Therefore, it is unfair to compare farm-raised fish with wild fish regardless of its nutritional merits.

The last but not the least, fish farming increases the eatable protein amount as pose to wastefulness. It is true that some fish farms do use fish meal, however, these fish meals are actually made of ineatable fish. By fish farming, people can in fact make extra eatable fish sources.

Independent Essay (20073264; 4 points)

In today's human being society, the value of cooperation has been weighed manifestly more than it was ever before. Despite its own merits, cooperation is required both by the changing environment and the developing technology. With all aspects carefully considered, i would agree that the ability to cooperate well with others is far more important than it was in the past.

First and for most, the change of human beings' life style promotes the importance of cooperation. Ere long, a family can live in an isolated or partly isolated life by planting their own food and make their own clothes. On the contrary, it is definatly unrealistic in today's modern life society, this is to say an ability to cooperate is essential for the living purpose. The needs to exchange food, clothes, knowledge etc all require cooperation. A person can not live well without cooperation skills, which is the fundamental requirement of the modern environment.

In addition, the continuous developing technology, provides an overwhelming amount of information in each and every occupation. In the past, a job may be done properly by an individual easily. However, today, there is totally a different story. Imagine how much work you will have with the simple job of doing a 1000 people research, the complicated analysis

programs, the time take for interview, the satistics discussion and the overall report. Is that a piece of cake?

The last point is today's society value teamwork much more than before. We are doing so because we are teaching so. In today's universities, colleges, high schools even primary schools, students are taught to cooperate with others well. A good team work ability is actually an basic goal of today's education.

Moreover, cooperation itself provides lots of benefits to a person. It can broaden him/her eyeview, make him/her thoughtful and reduce the stress.

Nowadays, cooperation is the vital ability required and learned by all kinds of people as it plays a more important role than any period in the past.

Coh-Metrix indices	Independent	Integrated
Number of sentences per paragraph	3.333	2.25
Number of syllables per word	1.721	1.42
Number of paragraphs per text	6	4
Number of sentences per text	20	9
Number of words per text	340	212
Syntactic similarity (sentence to sentence adjacent)	0.120	0.119
Syntactic similarity (sentence to Sentence within paragraph)	0.113	0.135
Syntactic similarity (sentence to sentence)	0.122	0.163
Number of higher-level constituents per word	0.676	0.736
Number of words before the main verb	4.150	2.667
Number of modifiers per noun phrase	1.023	0.911
Ratio of causal particles to causal verbs	0.455	0.500
Causal verbs	29.412	14.151
Number of causal verbs and particles	44.118	23.585
Positive causal connectives	14.706	9.434
Word concreteness (all words)	297.134	318.641
Word concreteness (content words)	357.172	418.971
Word familiarity (content words)	579.777	579.197
Word familiarity (all words)	593.387	597.524
Word Imagability (content words)	392.192	437.085
Word Imagability (all words)	324.295	328.387
Word meaningfulness (content words)	432.022	415.962
Word meaningfulness (all words)	347.264	325.947
Semantic similarity (LSA sentence to sentence adjacent)	0.175	0.256
Given/new information (LSA)	0.268	0.287
Semantic similarity (LSA sentence to sentence)	0.180	0.220
Nominalizations	55.882	18.868
All connectives	82.353	66.038
Positive logical connectives	23.529	28.302

Conditional connectives	0	4.717
Logical operators	26.471	37.736
Aspect repetition	0.842	1
Tense and aspect repetition	0.658	1
Tense repetition	0.316	1
Tense repetition	0.474	1
Lexical diversity (D)	109	43
Lexical diversity (M)	0.018	0.026
Lexical diversity (McCarthy score)	115.464	38.401
Lexical diversity (Vocd)	98.677	40.935
Word hypernymy	1.698	2.082
Hypernymy values of nouns	5.936	6.470
Hypernymy values of verbs	1.518	1.566
Word polysemy	3.607	4.294
Argument overlap (binary maximum user specified sentences unweighted)	0.255	0.694
Stem overlap (binary maximum user specified sentences unweighted)	0.324	0.75
Argument overlap (binary adjacent sentences unweighted)	0.316	0.875
Content word overlap (proportional adjacent sentences unweighted)	0.026	0.14
Content word overlap (proportional next 2 sentences unweighted)	0.041	0.152
Content word overlap (proportional next 3 sentences unweighted)	0.057	0.145
Noun overlap (binary adjacent sentences unweighted)	0.263	0.875
Noun overlap (binary next 2 sentences unweighted)	0.297	0.733
CELEX word frequency (content words minimum in sentence)	1.318	0.880
CELEX word frequency (content words written frequency in sentence)	0.352	0.614
CELEX word frequency (content words in sentence)	2.342	2.217
CELEX word frequency (content words)	2.420	2.330
CELEX word frequency (all words in sentence)	2.675	3.032
CELEX word frequency (all words)	3.096	3.203
Noun (singular or mass, POSnn)	214.706	207.547
Noun (plural, POSnns)	61.765	61.321
Prepositional phrase (POSpp)	105.882	113.208
Personal pronoun (POSprp)	35.294	42.453
Personal pronoun possessive case (POSprps)	14.706	4.717
Embedded clause (POSsbar)	26.471	99.057
Verbs in base form (POSvb)	41.176	28.302
Verbs in past tense (POSvbd)	5.882	9.434
Gerund or present participle verbs (POSvbg)	20.588	14.151
Past participle verbs (POSvbn)	29.412	23.585
Verbs in non-3rd person singular present form (POSvbp)	2.941	23.585
Verbs in 3rd person singular present form (POSvbz)	41.176	66.038
Verb phrases (POSvp)	182.353	221.698

APPENDIX M

ANOVA RESULTS OF ALL THE COH-METRIX INDICE

Means (standard deviations), F Values, and Effect Sizes for All the Indices in the 2007 Total Set

Coh-Metrix Indices	Independent	Integrated	F (1, 478)	p	η^2
Word concreteness (content words)	347.302 (15.031)	414.875 (18.837)	2362.294	.000	.908
Word Imagability (content words)	380.479 (14.518)	438.202 (18.218)	1920.002	.000	.889
Word concreteness (all words)	294.536 (8.502)	326.520 (13.275)	1251.591	.000	.840
Word familiarity (content words)	583.163 (6.491)	569.463 (4.816)	1044.069	.000	.814
CELEX word frequency (content words)	2.604 (0.146)	2.348 (0.126)	974.549	.000	.803
Word familiarity (all words)	596.726 (3.378)	588.650 (3.498)	972.182	.000	.803
CELEX word frequency (content words in sentence)	2.560 (0.150)	2.303 (0.142)	913.141	.000	.793
Number of words per text	312.370 (77.457)	197.130 (50.834)	860.109	.000	.783
Nouns (singular or mass; POSnn)	152.930 (31.759)	228.367 (35.940)	754.482	.000	.759
Word imagability (all words)	320.663 (8.925)	345.387 (13.509)	743.464	.000	.757
Word hypernymy	1.395 (0.208)	1.841 (0.232)	646.368	.000	.730
Stem overlap (binary maximum user specified sentences unweighted)	0.402 (0.187)	0.764 (0.187)	553.875	.000	.699
Noun overlap (binary next 2 sentences unweighted)	0.333 (0.177)	0.714 (0.206)	537.465	.000	.692
Number of sentences per text	16.630 (6.018)	9.870 (3.255)	468.633	.000	.662
Argument overlap (binary maximum user specified sentences unweighted)	0.437 (0.176)	0.749 (0.187)	464.648	.000	.660
Nominalizations	11.360 (6.166)	3.600 (2.398)	455.889	.000	.656
Noun overlap (binary adjacent sentences unweighted)	0.357 (0.189)	0.727 (0.215)	433.879	.000	.645

Verbs in base form (POSvb)	52.902 (16.770)	28.250 (15.882)	422.745	.000	.639
CELEX word frequency (all words)	3.227 (0.093)	3.094 (0.114)	397.899	.000	.625
Number of higher-level constituents per word	0.766 (0.037)	0.711 (0.036)	388.725	.000	.619
Lexical diversity (Vocd)	77.411 (17.422)	52.535 (140412)	378.055	.000	.613
Number of modifiers per noun phrase	0.738 (0.162)	0.966 (0.170)	340.140	.000	.587
Verbs in 3rd person singular present form (POSvbz)	26.362 (11.328)	48.873 (19.045)	308.548	.000	.564
Semantic similarity (LSA sentence to sentence adjacent)	0.176 (0.061)	0.279 (0.079)	305.399	.000	.561
Personal pronouns (POSprp)	55.185 (24.478)	26.778 (17.283)	299.352	.000	.556
Content word overlap (proportional next 3 sentences unweighted)	0.101 (0.040)	0.167 (0.055)	275.450	.000	.535
Content word overlap (proportional next 2 sentences unweighted)	0.105 (0.043)	0.172 (0.056)	261.182	.000	.522
Argument overlap (binary adjacent sentences unweighted)	0.526 (0.185)	0.785 (0.191)	258.512	.000	.520
Lexical diversity (McCarthy score)	74.635 (16.585)	55.164 (14.281)	254.652	.000	.516
Personal pronoun possessive case (POSprps)	15.378 (10.184)	3.799 (5.437)	245.773	.000	.507
Semantic similarity (LSA sentence to sentence)	0.161 (0.065)	0.273 (0.100)	244.187	.000	.505
Verb phrases (POSvp)	236.124 (32.205)	197.877 (29.632)	231.241	.000	.492
Content word overlap (proportional adjacent sentences unweighted)	0.113 (0.046)	0.179 (0.060)	220.462	.000	.480
Past participle verbs (POSvbn)	13.174 (9.668)	26.966 (15.635)	211.484	.000	.469
Verbs in non-3rd person singular present form (POSvbp)	36.902 (15.936)	22.484 (13.002)	175.379	.000	.423
Number of syllables per word	1.555 (0.109)	1.479 (0.079)	164.460	.000	.408
Hypernymy values of nouns	5.464 (0.558)	5.978 (0.542)	156.551	.000	.396

CELEX word frequency (all words in sentence)	2.872 (0.138)	2.769 (0.150)	145.939	.000	.379
Logical operators	45.126 (15.841)	33.824 (13.594)	94.132	.000	.283
Syntactic similarity (sentence to sentence)	0.093 (0.028)	0.112 (0.038)	91.411	.000	.277
Number of paragraphs per text	4.830 (1.835)	3.830 (1.607)	90.111	.000	.274
Lexical diversity (D)	72.820 (24.003)	58.730 (16.081)	87.990	.000	.269
Conditional connectives	4.295 (4.965)	1.102 (2.838)	80.774	.000	.253
Hypernymy values of verbs	1.317 (0.167)	1.417 (0.233)	76.738	.000	.243
Word meaningfulness (content words)	423.541 (14.607)	433.820 (14.037)	72.175	.000	.232
Verbs in past tense (POSvbd)	14.963 (10.954)	7.004 (9763)	71.236	.000	.230
Causal verbs	23.406 (10.045)	16.816 (9.848)	62.928	.000	.208
CELEX word frequency (content words average minimum in sentence)	1.348 (0.226)	1.230 (0.225)	61.017	.000	.203
Ratio of causal particles to causal verbs	0.763 (0.608)	1.353 (1.194)	54.193	.000	.185
Noun (plural, POSnns)	71.451 (22.194)	59.516 (21.537)	43.775	.000	.155
Positive causal connectives	16.908 (9.466)	22.093 (11.328)	38.437	.000	.139
Word polysemy	3.945 (0.420)	3.731 (0.463)	32.888	.000	.121
Syntactic similarity (sentence to sentence within paragraph)	0.098 (0.034)	0.112 (0.047)	32.166	.000	.119
Tense aspect repetition	0.581 (0.170)	0.700 (0.283)	31.964	.000	.118
Number of sentences per paragraph	3.900 (2.294)	3.115 (2.048)	29.996	.000	.112
Positive logical connectives	34.008 (12.086)	39.934 (15.746)	28.446	.000	.106
Tense repetition	0.641 (0.164)	0.752 (0.275)	28.006	.000	.105

Syntactic similarity (sentence to sentence adjacent)	0.098 (0.033)	0.109 (0.042)	23.919	.000	.091
Tense and aspect repetition	0.769 (0.104)	0.839 (0.218)	20.196	.000	.078
Gerund or present participle verbs (POSvbg)	13.702 (9.702)	17.124 (10.228)	18.076	.000	.070
Given/new information (LSA)	0.296 (0.037)	0.310 (0.049)	17.393	.000	.068
Lexical diversity (M)	0.022 (0.003)	0.023 (0.003)	15.834	.000	.062
Word meaningfulness (all words)	352.836 (10.927)	349.752 (11.476)	12.914	.000	.051
CELEX word frequency (content words written frequency in sentence)	0.304 (0.088)	0.279 (0.098)	9.772	.002	.039
Embedded clauses (POSsbar)	51.355 (16.526)	54.981 (17.225)	7.778	.006	.032
Prepositional phrases (POSpp)	112.084 (22.447)	116.447 (25.129)	5.187	.024	.021
Aspect repetition	0.897 (0.120)	0.925 (0.224)	3.434	.065	.014
Number of words before the main verb	5.020 (1.938)	4.784 (2.142)	2.048	.154	.008
All connectives	85.359 (18.104)	87.419 (20.810)	1.563	.212	.006
Number of causal verbs and particles	41.017 (14.603)	39.986 (15.163)	.657	.418	.003

APPENDIX N

ANOVA RESULTS OF THE INTEGRATED ESSAYS

F Values and p Values for the Integrated Essays (Undergraduate vs. Graduate Applicants)

	F	p
Number of sentences per paragraph	6.548	.012
Personal pronoun (POSprp)	5.717	.019
Number of paragraphs per text	4.714	.032
Lexical diversity (McCarthy score)	3.691	.058
Normalizations	2.755	.100
Lexical diversity (D)	2.589	.111
Positive causal connectives	2.501	.117
Prepositional phrase (POSpp)	2.414	.124
causal verbs and particles	2.406	.124
CELEX word frequency (context words minimum in sentence)	2.129	.148
Content word overlap (Proportional adjacent sentences unweighted)	2.122	.148
Number of Syllables per word	2.045	.156
Lexical diversity (M)	2.012	.159
CELEX word frequency(content words written frequency in sentence)	1.784	.185
Personal pronoun possessive case (POSprps)	1.601	.209
Syntax similarity (sentence to sentence within paragraph)	1.541	.217
Content word overlap (Proportional next 2 sentences unweighted)	1.400	.240
Number of words per text	1.379	.243
Verbs in past tense (POSvbd)	1.238	.269
Mean number of words before the main verb	1.149	.286
Content word overlap (Proportional next 3 sentences unweighted)	1.120	.293
Syntax similarity (sentence to sentence adjacent)	.951	.332
Semantic similarity (LSA sentence to sentence)	.935	.336
Word polysemy	.813	.369
Conditional connectives	.724	.397
Word meaningfulness (all words)	.681	.411
Noun (singular or mass, POSnn)	.650	.422
Noun overlap (Binary adjacent sentences unweighted)	.592	.444
Semantic similarity (LSA sentence to sentence adjacent)	.566	.453
Word familiarity (content words)	.552	.459
CELEX word frequency (all words in sentence)	.547	.461
Syntax similarity (sentence to sentence)	.544	.462
Word concreteness (all words)	.540	.464
Verb phrases (POSvp)	.537	.466
Word imagability (all words)	.502	.480
Lexical diversity (vocd)	.488	.487
Argument overlap (Binary adjacent sentences unweighted)	.474	.493

CELEX word frequency (content words)	.449	.504
CELEX word frequency (all words)	.419	.519
Hypernymy values of nouns	.407	.525
Stem overlap (Binary maximum user specified sentences unweighted)	.396	.530
Positive logical connectives	.352	.554
Number of modifiers per noun phrase	.333	.565
Tense and aspect repetition	.322	.572
Gerund or present participle verbs (POSvbg)	.311	.578
Verbs in non-3rd person singular present form (POSvbp)	.301	.584
Aspect repetition	.299	.586
Hypernymy values of verbs	.277	.600
Verbs in 3rd person singular present form (POSvbz)	.248	.619
Causal verbs	.246	.621
Noun overlap (Binary next 2 sentences unweighted)	.244	.622
Argument overlap (Binary maximum user specified sentences unweighted)	.191	.663
Number of higher-level constituents per word	.179	.673
CELEX word frequency (content words in sentence)	.158	.692
Logical operators	.156	.693
Verbs in base form (POSvb)	.156	.694
All connectives	.145	.704
Word familiarity (all words)	.116	.735
Given/new information (LSA)	.108	.743
Word hypernymy	.102	.750
tense repetition	.094	.759
Embedded clause (POSsbar)	.051	.821
Noun (singular or mass, POSnn)	.048	.828
Ratio of causal particles to causal verbs	.024	.877
Word imagability (content words)	.021	.885
Word concreteness (content words)	.014	.906
Number of sentences per text	.007	.932
Past participle verbs (POSvbn)	.004	.948
Noun (plural, POSnns)	.001	.979
Word meaningfuness (content words)	.001	.980

APPENDIX O

ANOVA RESULTS OF THE INDEPENDENT ESSAYS

F Values and p Values for the Independent Essays (Undergraduate vs. Graduate Applicants)

	F	p
Hypernymy values of verbs	9.312	.003
Verb phrases (POSvp)	6.759	.011
Gerund or present participle verbs (POSvbg)	4.637	.034
Noun (plural, POSnns)	4.604	.034
Embedded clause (POSsbar)	4.388	.039
Prepositional phrase (POSpp)	3.469	.066
Mean number of words before the main verb	3.300	.072
Positive causal connectives	3.187	.077
CELEX word frequency (content words)	2.906	.091
Verbs in past tense (POSvbd)	2.800	.097
causal verbs and particles	2.681	.105
Word polysemy	2.651	.107
Ratio of causal particles to causal verbs	2.638	.108
Word hypernymy	2.333	.130
Aspect repetition	2.304	.132
Number of sentences per text	2.264	.136
Noun (singular or mass, POSnn)	2.194	.142
CELEX word frequency (context words minimum in sentence)	2.193	.142
Word concreteness (content words)	1.898	.171
CELEX word frequency (all words)	1.634	.204
CELEX word frequency (content words in sentence)	1.437	.233
Argument overlap (Binary adjacent sentences unweighted)	1.420	.236
Hypernymy values of nouns	1.415	.237
tense repetition	1.414	.237
Word concreteness (all words)	1.331	.251
Number of higher-level constituents per word	1.285	.260
Content word overlap (Proportional next 2 sentences unweighted)	1.261	.264
Content word overlap Proportional next 3 sentences unweighted	1.257	.265
CELEX word frequency (all words in sentence)	1.216	.273
Argument overlap (Binary maximum user specified sentences unweighted)	1.191	.278
Positive logical connectives	.993	.322
Number of words per text	.878	.351
Verbs in 3rd person singular present form (POSvbz)	.845	.360
Given/new information (LSA)	.771	.382
CELEX word frequency(content words written frequency in sentence)	.759	.386

Logical operators	.724	.397
Word meaningfulness (content words)	.696	.406
Number of Syllables per word	.649	.423
Verbs in non-3rd person singular present form (POSvbp)	.563	.455
Semantic similarity (LSA sentence to sentence adjacent)	.481	.490
Word imagability (content words)	.466	.497
Causal verbs	.458	.500
Stem overlap (Binary maximum user specified sentences unweighted)	.380	.539
Content word overlap (Proportional adjacent sentences unweighted)	.362	.549
Word imagability (all words)	.339	.562
Personal pronoun (POSprp)	.338	.562
Word familiarity (content words)	.326	.569
Lexical diversity (M)	.290	.591
All connectives	.266	.607
Word familiarity (all words)	.209	.649
Conditional connectives	.163	.687
Syntax similarity (sentence to sentence adjacent)	.158	.692
Number of modifiers per noun phrase	.099	.753
Syntax similarity (sentence to sentence)	.093	.760
Noun overlap (Binary adjacent sentences unweighted)	.092	.763
Noun overlap (Binary next 2 sentences unweighted)	.089	.766
Noun (singular or mass, POSnn)	.089	.766
Lexical diversity (McCarthy score)	.074	.786
Lexical diversity (D)	.074	.786
Number of sentences per paragraph	.062	.804
Verbs in base form (POSvb)	.047	.828
Past participle verbs (POSvbn)	.033	.857
Lexical diversity (vocd)	.028	.867
Semantic similarity (LSA sentence to sentence)	.019	.889
Tense and aspect repetition	.007	.935
Personal pronoun possessive case (POSprps)	.006	.939
Syntax similarity (sentence to sentence within paragraph)	.003	.956
Normalizations	.002	.968
Number of paragraphs per text	.000	.997
Word meaningfulness (all words)	.000	.999