An Empirical Study on Predictability of Writing Form to Second Language Writing Performance —Analysis Based on SEM Approach

By

Gao Yuan

Supervisor: Prof. Gao Guixiang

Specialty: English Language and Literature

Research Direction: English Linguistics

Submitted to

College of Foreign Languages

In Partial Fulfillment of the Requirements

For the Degree of Master of Arts

At Hebei Normal University

September 19, 2010

Y1816330

学位论文原创性声明

本人所提交的学位论文“An Empirical Study on Predictability of Writing Form to Second Language Writing Performance—Analysis Based on SEM Approach”，是在导师的指导下，独立进行研究工作所取得的原创性成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出重要贡献的个人和集体，均已在文中标明。

本声明的法律后果由本人承担。

论文作者（签名）：高媛　　指导教师确认（签名）：

2010年12月6日　　年　月　日

学位论文版权使用授权书

本学位论文作者完全了解河北师范大学有权保留并向国家有关部门或机构送交学位论文的复印件和磁盘，允许论文被查阅和借阅。本人授权河北师范大学可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或其它复制手段保存、汇编学位论文。

（保密的学位论文在_____年解密后适用本授权书）

论文作者（签名）：高媛　　指导教师（签名）：

2010年12月6日　　年　月　日

Abstract

Holistic or analytic scoring has been an eternal topic in the field of testing in recent years. Holistic scoring is the assigning of a single score to a script based on the overall judgment about its quality. In terms of analytic scoring, scripts are rated on several aspects of scoring rubric rather than a single score.

The study is to construct the model on predictability of writing form to writing performance. 243 English majors participated in the study. They have been required to write the composition in 50 minutes during the final examination. The students come from Grade Three of different universities in Hebei Province. Namely, Hebei University, Hebei University of Technology and Hebei University of Economics and Business. All of the compositions need to be scored by three raters twice, one with holistic scoring and the other with analytic scoring. Holistic scoring is based on TEM4, whereas analytic scoring adopts the criteria: relevance、explicitness、coherence、sufficiency (Wen, 2007); fluency、accuracy and complexity (Wolfe-Quintero et al.，1998; Liang 2005). After rating, all of the data should be analyzed by SPSS (16.0). AMOS (7.0) is used to establish SEM based on literature and hypothesis. In the study, writing form is indicated by three observed variables: fluency, accuracy and complexity. Writing performance is seen as latent variable. SEM is used to construct and check to what extent these three variables can predict writing performance.

The objective of the paper is to provide detailed analysis about writing form and writing performance. The paper attempts to address the following questions: in L2 writing, what is the relationship between writing form and writing performance? As for the second language learners, can fluency、accuracy and complexity predict writing performance? As for high-proficiency learners and low-proficiency learners, how do different understanding of writing form influence writing performance?

According to data analysis and interviewees' response, there are three major findings in the study. First, writing form can predict 47.61% variance of writing performance directly. Second, fluency、complexity and accuracy are positively correlated with writing performance.

Complexity and fluency can predict more about writing performance than accuracy. Third, different understanding of writing form can affect writing performance directly according to interviewees' response. In sum, the findings are helpful for us to improve the way of assessing L2 writing performance. Meanwhile, fluency、complexity and accuracy can provide more effective feedback to the test participants from different multi-facet, which is helpful to improve the writing performance.

Key Words: writing form; writing performance; predictability; SEM

摘 要

近年来，二语写作成绩评估成为热点话题之一。整体评估法是指评分者快速通读作文，依据特定的评分量表评分。而分析评估法则是评分者依据评分细则给不同的项目分别打出分数。

本研究旨在提出二语写作形式对二语写作成绩的预测力的理论模型。参加研究的243名被试为河北省三所高等院校英语专业三年级学生，写作任务为命题限时作文。评分员6名，分为两组，其中3名评分员以英语专业四级考试为评估指标进行整体性评分，另3名评分员进行分析性评分，采用作文切题性、论点明确性、篇章连贯性、说理透彻性（文秋芳，2007），流利性、准确性和复杂性（Wolfe-Quintero et al., 1998; 梁茂成，2005）这7个指标评分。评分结束后，本文通过SPSS（16.0）进行数据分析，以AMOS(7.0)为工具，基于文献理论与相关假设，确定“写作成绩”为内生潜在变量，“写作形式”为外生潜在变量，流利性，准确性和复杂性为观测变量，构建并验证写作形式对写作成绩预测力的结构方程模型。

本研究探讨并期望解决在二语写作评估中，写作形式与写作成绩相关度如何；反映写作形式的三个指标即流利性、准确性和复杂性能否在一定程度上预测写作成绩；以及通过访谈形式探讨高分组和低分组被试对流利性，准确性和复杂性的理解是否影响和预测写作成绩。

根据数据分析以及对被试访谈，本研究有三个主要发现。其一，写作形式可以预测二语写作成绩47.61%的方差。其二，流利性、准确性和复杂性与写作成绩之间呈正相关。复杂性和流利性对二语写作成绩有较强的预测力，准确性对于写作成绩预测力较弱。其三，不同成绩的学习者对于流利性、准确性和复杂性的理解直接影响二语写作成绩。本研究有助于改进二语写作评估方法，同时以流利性、准确性和复杂性为评分标准的分项评分法能够对学生作文提供多种层面的有效反馈，具有较强的教学启发意义。

关键词： 写作形式; 写作成绩; 预测力; 结构方程模型

List of Abbreviations

AES	Automated Essay Scoring
AGFI	Adjusted Goodness of Fit Index
AMOS	Analysis of Moment Structures
ANOVA	Analysis of Variance
CET 4	College English Test Band 4
CET 6	College English Test Band 6
CFI	Comparative Fit Index
Df	Degrees of freedom
EFL	English as a Foreign Language
GFI	Goodness of Fit Index
L1	First Language
L2	Second Language
LISREL	Linear Structural Equation Mode
RMSEA	Root Mean Square Error of Approximation
SEM	Structural Equation Model
SLA	Second Language Acquisition
SPSS	Statistical Package for Social Sciences
TEEP	Test in English for Educational Purposes
TEM 4	Test of English for English Majors Band 4
TEM 8	Test of English for English Majors Band 4
TOEFL	Test of English as a Foreign Language

List of Tables

Table 2.1 Measures of Writing Quality in Previous Study.....14

Table 2.2 A Comparison of Holistic and Analytic Scales.....18

Table 3.1 Demographic Information of Subjects.....25

Table 4.1 Correlations among Raters in the Holistic Scorings and Analytic Scoring.....32

Table 4.2 Correlations among Raters in the Holistic Scorings.....33

Table 4.3 Mean and Standard Deviation of Holistic Scoring.....34

Table 4.4 Correlations among Raters on Fluency in the Analytic Scoring.....34

Table 4.5 Mean and Standard Deviation of Fluency Scores.....35

Table 4.6 Correlations among Raters on Complexity in the Analytic Scoring.....35

Table 4.7 Mean and Standard Deviation of Complexity Scores.....36

Table 4.8 Correlations among Raters on Accuracy in the Analytic Scoring.....36

Table 4.9 Mean and Standard Deviation of Accuracy Scores.....37

Table 4.10 Model Fit Summary.....41

Table 4.11 Holistic Score of Writing Performance of Subjects.....46

Table 4.12 Analytic Score of Form of Subjects.....46

List of Figures

Figure 2.1 Measuring the Validity of the Construct "Essay Content" by Using SEM....23

Figure 4.1 Preliminary Model.....38

Figure 4.2 Refined Model 1.....39

Figure 4.3 Refined Model 2.....40

Contents

Abstract IV

摘 要 VI

List of Abbreviations VII

List of Tables VIII

List of Figures IX

Chapter One Introduction 1

1.1 Research Background 1

1.2 Purpose of the Study 3

1.3 Significance of the Study 4

1.4 Organization of the Paper 4

Chapter Two Literature Review 6

2.1 Understanding Writing Form 6

2.1.1 Definition and Measures of Fluency 6

2.1.2 Definition and Measures of Accuracy 8

2.1.3 Definition and Measures of Complexity 10

2.2 Understanding Writing Performance 13

2.2.1 Different Criteria of Writing Performance 13

2.2.2 Writing Rating Methods 15

2.3 Summary of Studies on Writing Form and Writing Performance 19

2.4 Empirical Studies on Structural Equation Modeling 21

Chapter Three Methodology 24

3.1 Research Questions 24

3.2 Subjects 24

3.3 Instruments 26

3.3.1 Essay Writing 26

3.3.2 Interview 27

3.4 Data Collection 27
3.5 Rating 28
3.5.1 Holistic and Analytic Scoring Scales 28
3.5.2 Rating Procedure 29
Chapter Four Results and Discussion 32
4.1 Data Presentation 32
4.1.1 Score Reliability 32
4.1.2 Correlation Analysis 33
4.1.3 SEM Analysis 37
4.2 Discussions 42
4.2.1 Discussion on Research Question 1 42
4.2.2 Discussion on Research Question 2 44
4.2.3 Discussion on Research Question 3 45
Chapter Five Conclusion 51
5.1 Major Findings 51
5.2 Implications 52
5.3 Limitations 54
5.4 Suggestions for Future Research 55
Appendix Ⅰ 二语写作评估标准（内容层面） 65
Appendix Ⅱ 二语写作评估标准（形式层面） 67
Appendix Ⅲ TEM4 作文评分标准 68
Appendix Ⅳ Some Compositions for Analysis 69
Acknowledgments 74
Academic Achievements 75

Chapter One Introduction

1.1 Research Background

L2 writing (Second Language Writing), as one of the most difficult skill for learners to develop, has aroused more attention by researchers aboard and in China in recent years.

Most of the scholars and experts abroad focused their attention on the following fields: writing theories; writing process, including writing strategies, differences of individual learners; writing results, including text analysis, error analysis and corpus analysis; factors affecting writing, focusing on register knowledge, motivation and demand; writing teaching method; writing assessment; curriculum design and so forth. (Atkinson 2003; Matsuda 2003)

Recently, most specialists in China focused their attention on writing theories, writing process, text interactions, writing factors and teaching writing. It is not difficult to arrive at the conclusion that from 1993 to 2004, empirical research about L2 writing covered four aspects: (1) How does L1 perform in L2 writing of Chinese learners? (2) What kind of factors of language and context appear? (3) What factors can affect L2 writing ability? (4) How can improve and better L2 writing teaching in China? (Wang Lifei, 2005).

As one field of L2 writing, L2 writing assessment has caused much attention in recent years.

When it comes to writing assessment, there is no doubt that there remain a lot of controversies in scoring or rating. Two main methods of assessing writing, holistic scoring and analytic scoring, have caused great study in western countries. Most scholars in western countries focused more attention on comparison of different rubrics: holistic and analytic scoring. Researchers have discussed a lot about whether a single score should be given to each composition or different dimensions should be scored separately. At the same time, they have compared validity and reliability of two methods of assessment. Validity and reliability are central to effective testing practice. Validity, or judgments of the relevance of test tasks to the intended score representation, is thus at the heart of the performance test (Kenyon, 1998). Previous studies on writing assessment aimed to investigating factors affecting scoring,

including variables of task、rater、text、context and test-taker.

Another aspect of writing assessment is about the criteria. Different criteria had been proved to be useful and effective to test learners' achievement in L2 writing. Jacobs et al (1981) stated content, organization, vocabulary, language use and mechanics can be regarded as effective means of assessment criteria. Purves (1985) made a summary about the major criteria: content, organization, style, conventions, and communication requirements. Weir (1990) made a careful classification about the criteria. There were seven dimensions in the criteria: relevance and adequacy of content, compositional organization, cohesion, adequacy of vocabulary for purpose, grammar, mechanical accuracy I (punctuation) and mechanical accuracy II (spelling). Hamp-Lyons (1990) viewed ideas and argument, rhetorical features, and language control as major criteria. Blok and de Glopper (1992) states content, organization, style and conventions can be applied in testing. In China, Liang (2005) pointed out that language, content, and organization can become major constructs of assessing EFL writing. Wen (2007) believed the criteria established by Hamp-Lyons was not suitable to L2 writing due to its simplicity. It was not easy to make clear distinction among dimensions in the criteria which established by Jacobs et al (1981) and Weir (1990). She used form and content to analyze L2 scoring. Vocabulary, sentence pattern, grammar, and mechanics can cover aspects of form while relevance, explicitness, coherence, sufficiency were related to content.

In China, L2 writing assessment have adopted holistic scoring either in theory or in practice in most time, ranging from classroom practicing to large-scaled tests, such as CET4、CET6、TEM4 and TEM8. As for analytic scoring, some researchers have probed into this field. Liang (2005) used analytic scoring to conduct empirical study to build scoring system and Wen (2007) employed it to validate content construct.

Empirical research relating testing of L2 writing in China is not adequate. Liang (2005) discussed the predictability from content, language and organization in order to establish computer-assisted scoring system. If a computer-assisted scoring system can be proved useful and applied into practice, it can reduce tremendous amount of work of raters especially in large-scaled test. However, content and form can cover the assessment of L2 writing. As for

structure, it can be included by content and form according to its features. Wen (2007) established the model by using Structure Equation Modeling (SEM) to measure the validity of the construct "essay content". Although some researchers probed into L2 scoring from different aspects, there is no agreement in scoring approach and scoring criteria.

1.2 Purpose of the Study

This paper attempts to explore the contribution of form to writing performance of time-limited composition from English major students in university. Form can be conveyed by fluency, accuracy, and complexity, which are sufficient perform of writing performance. The paper aims to test the correlation of the three dimensions and the contribution of them to form, as well as form to writing performance.

The major purpose of the study is to investigate how writing form predict writing performance in L2 writing. The study hopes to state clearly different predictability of dimensions to writing performance.

The study comprises two parts—the quantitative part and the qualitative part. The quantitative part of the study attempts to prove into some features about form of learners' writing, especially the application of three dimensions—fluency, accuracy and complexity. These three dimensions were defined by Wolfe-Quintero et al. (1998). And these three dimensions have been widely used to test oral English as well as written English. In the quantitative part, SEM is established to illustrate the predictability of each dimension to writing performance. SEM, as one effective tool, can provide objective and efficient result to the study.

The qualitative part of the study is designed to prove how different level of students behaved and performed in L2 writing, especially in time-limited composition. Interview is used as research method. According to interviewees' responses, it is clearly to see different understanding of form affecting high proficiency level students and low proficiency level students.

Anyway, the conclusion reached from research questions might be helpful to both teachers and learners.

1.3 Significance of the Study

The major findings of the empirical research may play significant role in the field of L2 writing.

From methodological perspective, SEM has been used to establish model which can give strong and powerful argument to illustrate the relationship between form and writing performance. In addition, predictability of different dimensions of form to writing performance can be seen clearly from the model. SEM can be applied in foreign language research, because it has lots of advantages.

From pedagogical perspective, both teachers and students should clarify the relationship between form and writing performance. Traditionally, holistic scoring has been widely used in the tests and classroom teaching, ignoring the application of analytic scoring. The empirical findings have provided more information about two different scoring. The advantages of analytic scoring can be seen clearly. Teachers should lay emphasis on analytic scoring in teaching writing, giving more detailed feedback about content and form to students. In addition, teachers should emphasize both form and content, which has been one effective way of improving students' L2 writing performance.

1.4 Organization of the Paper

This paper consists of five chapters.

Chapter One provides the purpose and the background of the study. It is composed of research background, purpose of the study, significance of the study and organization of the paper.

Chapter Two mainly reviews the relevant literature on the present study. One part is about definition and measurement of writing form and writing performance. Another part is about two methods about assessing writing performance, both of which are adopted by the study. Holistic scoring and analytic scoring had been proved by scholars in western countries and in China. At last, brief review of related study about SEM applied in language testing research is introduced.

Chapter Three describes the design of the present study, including objective of the study,

subjects, instrument, data collection, rating scales, rating training and rating procedure.

Chapter Four is results and discussion. In this chapter, it presents data analysis as well as discussions of three research questions. On the one hand, data analysis reports score reliability of human raters from inter- and intra- reliability. Correlation analysis is conducted by SPSS (16.0) to reveal the reliability of scores and compare the results of holistic scoring and analytic scoring. AMOS (7.0) is used to construct SEM concerning the topic. On the other hand, three research questions are discussed respectively and thoroughly in this chapter.

The last chapter draws the conclusion of the major findings, discusses the implications, limitations and provides suggestions for future study of the topic.

Chapter Two Literature Review

This chapter probes into the literature concerning this study. It is composed of the following parts: understanding writing form, especially fluency, accuracy, and complexity, among which three dimensions can reflect writing form in the study; understanding assessing writing performance, holistic scoring and analytic scoring introduced thoroughly; previous studies on writing form and writing performance; and empirical studies on SEM.

2.1 Understanding Writing Form

Generally speaking, a piece of writing is a unit of content and form. Language usage is the main performance of form. According to the classification of some researchers, fluency, accuracy and complexity, these three dimensions are in the constructs of L2 proficiency, which can be applied to assess oral English as well as written English. (Larsen-Freeman and Strom, 1977, 1978; Larsen-Freeman, 1983; Foster and Skehan, 1996; Wolfe-Quintero et al, 1998).

In L2 writing assessment, how to judge the performance of language have aroused researchers' concern in the past years. As is known to all, most measures of assessing writing performance only focus on one aspect of essays: the quality of language. How to assess the performance of language has caused much attention in the past years. In this study, three categories had been discussed in order to assess the quality of language. It involves fluency, accuracy, and complexity. According to the literature review, fluency, accuracy and complexity have been regarded as primary variables in applied linguistic research.

2.1.1 Definition and Measures of Fluency

It is difficult to define the term fluency because of its diverse range of definitions. Some attempts have been made to fluency in order to find out what extent writing performance can be predicted by fluency.

Fluency is originally one item used in oral discourse, referring "the processing of language in real time" (Schmidt, 1992). Lesson (1975) defined fluency as "the ability of the speaker to produce indefinite many sentences conforming to the phonological, syntactical and

semantic exigencies of a given natural language on the basis of a finite exposure to a finite corpus of that language."

Foster and Skehan (1996), Larsen-Freeman and Strom (1977,1978), Wolfe-Quintero et al. (1998) viewed fluency as an independent measure together with accuracy and complexity.

However, Wolfe-Quintero et al. (1998) maintained that fluency can be used in the field L2 writing. They believe fluency measures can reveal how proficient the writer is to produce writing.

Wolfe-Quintero et al. (1998) argued that "fluency means that more words and more sentences are accessed in a limited time, whereas a lack of fluency means that only a few words or structures are accessed." In other words, fluency is one aspect of ability of subjects producing words and sentences as more as possible in limited time.

It is common to calculate the total number of words in a written article to measure fluency. Total number of scores has been one important aspect to assess writing performance and some experts had validated the results. In the previous study, Wolfe-Quintero et al. (1998) made the study about 39 studies to draw the conclusion that the best measure of fluency, should define the number of words per T-unit, the number of words per clause the number of clauses per T-unit and the number of words per error-free t-unit. Wolfe-Quintero et al. (1998) argued that in the past years several researchers have confirmed a closely correlation between the length of the article and writing proficiency. Wolfe-Quintero et al. (1998) reported that a rather high correlation is 0.65 or above it. In Page's (1968) study, the number of the words in one essay was one of the best correlates of writing performance. The correlating was reported as r=0.32. Slotnick (1972) conducted analysis by multiple regression, including the length of the essay as one of the observed variable. Although he did not publish the correlation coefficient, he had found the length was definitely a significant correlate of writing quality. McCurry and McCurry (1992) reached the same conclusion in the regression model, which includes 27 variables.

In a similar study, Reid and Findlay (1986) and Carlson (1988) confirmed the findings of mentioned above by using regression models, including essay length as an observed variable.

Except for the total numbers of the essay, another two measures were included in

Slotnick's (1972) and McCurry and McCurry's (1992) models, that is to say, the number of word types, and the number of sentences. Both of the two elements were regarded as the predictive of writing scores.

In Fillmore's (1979) study, he indicated that fluency was not only a quantitative aspect but also a qualitative aspect. The number of word types can reflect both the quantity and the quality of text.

As to the correlation between fluency measures and writing proficiency, several researches have provided it. Wolfe-Quintero et al. (1998) reported a significant correlation at 0.65 between fluency and writing quality. Vann (1979) found the relationship between fluency and TOFEL scores. Fluency measure can suggest a significant correlation with TOFEL scores.

2.1.2 Definition and Measures of Accuracy

Most researchers had accepted the definition of accuracy, which is "freedom from error." Foster & Skehan (1996) had stated. Later, Skehan & Foster (1997) pointed that accuracy was focused on how well language was produced to match the rules of the target language. Simply speaking, Accuracy is concerned with how well language is produced in relation to the rule system of the target language. It refers to the ability of using language to performing both in spoken English and written English.

In terms of assessing accuracy, many researchers have made attempts and followed different measures.

Wolfe-Quintero et al. (1998) made the conclusion that error-free T-unit, error-free T-units per T-unit and errors per T-unit were the best measures to testify accuracy. They were distinct measures from complexity. One disadvantage of the measures was uncertainty about the nature of the errors. It can not show clearly the seriousness of the error and the types of errors the writer made, such as lexical, syntax, spelling or style of the error. Polio (1997) pointed out that it was rather a subjective matter to make the clearness of error, because "defining an error may be problematic." This view had been proved by other experts. For example, Casanave (1994) put forward the similar viewpoint "it was difficult to determine whether the writer had made an error or not." Larsen-Freeman (2006) based on

Wolfe-Quintero et al.'s (1998) and Larsen-Freeman and Strom's (1977) regarded accuracy as the proportion of error-free T-units to T-units.

Reviewing the literature, it is not difficult to find out that the use of prepositions and the use of articles were two measures regarding measures of accuracy, especially measures of appropriateness. In sum, measures of accuracy rely on learners' errors.

Different researchers have adopted different methods to measure errors.

Hedgcock&Lefkowitz (1992) utilized scale adapted from Jacobs et al., (1981) to make comparison about feedback on essays. Among which three aspects grammar, vocabulary and mechanics were related to accuracy. Engber (1995) made the classification about lexical error. One was lexical choice, including unconcerned of words, mistaken of synonyms, mistaken of collocation、misuse of phrase. Another was lexical form, including: misuse of derivative words、mistaken of verb forms、and misspelling. Larsen-Freeman (1977,1978,1983), Wolfe-Quintero et al. (1998) adopted lexical, spelling, morph syntactic, even punctuation errors, as the references in determining whether a T-unit was error-free.

The dimensions by Wolfe-Quintero et al. (1998) are widely used in L2 writing research in China. According to the classification by Wolfe-Quintero et al., (1998), errors can include morphological errors, lexical errors and syntactical errors. They have their respective range. For example, morphological errors refer to errors in propositions, articles, plural forms, derivational morphemes and so on. Lexical errors refer to errors in spelling, abstract words phrases, collocation and modifiers. Syntactic errors refer to errors in embedding, complementation, coordination, word order, fragments, or the missing of constituents.

In conclusion, a large number of researchers have realized that errors do occur in adult second language data. Errors are closely related to accuracy.

Accuracy has caused widely attention in traditional teaching. Teachers noticed more about accuracy in the following aspects: spelling, word usage and grammar mistakes while assessing writing performance. Among the literature, correct spelling, punctuation, capitalization, and grammar and legible handwriting are included in conventions, which is an independent aspect of assessing. Frankly, this dimension can be encompassed by accuracy. The framework of Test in English for Educational Purposes (TEEP) consists of three accuracy

dimensions: grammar, mechanical accuracy; punctuation, and mechanical accuracy and spelling (Weir, 1998).

Yong-Won Lee, Claudia Gentile, Robert Kantor (2008) put forward the view that in automated essay scoring, accuracy is a major concern. Grammatical accuracy, usage accuracy, mechanical accuracy and linguistic accuracy should be identified in E-rater. Mechanical accuracy always play a minor role in holistic scoring with respect to AES, whereas it has strong correlation to human raters' judgment.

The major concern of accuracy is in relation to the rules of target language. It can reflect fluency and complexity to some degree. Although Stephen Isaacson (1996) held the view that when fluency was focused, misspellings, faulty punctuation were not considered, although accuracy was one of the important aspects to evaluate writing performance.

2.1.3 Definition and Measures of Complexity

Foster and Skehan's (1996) stated that complexity refers to the capacity to use more advanced language, with the possibility that such language may not be controlled so effectively.

According to the classification by Foster and Skehan's (1996) and Wolfe-Quintero et al.'s (1998), this review adopted two sub-categories: lexical complexity and syntactic complexity.

2.1.3.1 Lexical Complexity

One aspect of complexity is lexical complexity. Researchers have given deep insights into this field in the past years. Lots of researches have been able to attempt to lexical complexity and writing performance. In this regard, according to some researchers, there is closely relationship between writer's lexicon and writing performance.

Grobe's (1981) study came to the conclusion that vocabulary diversity was closely related to the quality of writing, "what teachers currently perceived as "good" writing is closely associated with vocabulary diversity ", which is concerned by teachers in the current teaching. Hyltenstam (1988) stated that it was crucial for effective writing whether there was a large range of lexicon in a piece of article. That is to say, a wide range of lexicon can decide the performance of writing to a certain degree.

There is no doubt that writing performance is closely related to the writer's lexicon. A tremendous amount of studies probe into measuring writing performance with lexical features extractable from articles. With respect to measures of complexity, there exist different ones.

One specific measure, word length, have been tested a significant correlation between average word length and writing performance. The findings can be confirmed by Page (1968), Slotnick (1972), Carlson (1988), and Berglund and Mason (2001).

Nominalizations can be regarded as another word length-based measure refers to the long word formed by suffixation. MaCurry and McCurry's (1992) regression model was considered having strong predictive power, though there was no report about the correlation between nominalizations and writing performance.

Wolfe-Quintero et al. (1998) chose type-token ratio of word type variation per total words along with type-token ratio of sophisticated word types (based on frequency lists) per total word types as the best measures of lexical variety. Type-token ratio can imply lexical richness of a writer. As a matter of fact, it can be affected by length of the text.

Larsen-Freeman (2006) based on Wolfe-Quintero et al.'s (1998) and Larsen-Freeman and Strom's (1977) studies viewed grammatical complexity as average number of clauses per t-unit.

A good writing must be of diversifying words. That is to say, various kinds of words should be available. Harley and King (1989) found.

As Hyltenstam (1988) argued that it was essential for effective communication to compose of a wide range of lexicon. Wolfe-Quintero et al. (1998) stated clearly how varied or complex the words or word types were necessary for the composition to be assessed. That is to say, lexical complexity focuses more attention on how sophisticated the words or word types are used rather than how many words are present in the article.

Laufer (1994), Laufer & Nation (1995) held the view that the richness of vocabulary turned out to be the biggest factor influencing the final scores.

Reviewing the literature, it is easy to find that lexical complexity always focused on the following points: whether there is various kinds of basic and advanced words are available and even can be accessed quickly. Measures of lexical complexity are focused on how

sophisticated the words used.

2.1.3.2 Syntactic Complexity

According to Foster and Skehan (1996), it is not difficult to find out syntactic complexity is another aspect of complexity, referring to "a greater variety of syntactic patterning"

Wolfe-Quintero et al. (1998) stated that the number of clauses per T-unit and the dependent clauses per clause were the best measures. At the same time, they pointed out that those who can use a wide variety of both basic sentence pattern and sophisticated structures, can be called a syntactically complex writer. On the contrary, a syntactically simple writer can use a narrow range of basic structures. According to the study, how sophisticated the production units are more concerned than how many production units are presented in the composition.

Ortega's (2003) put forward the definition of syntactic complexity, which was concerned with the different range of forms in language production as well as the extent of sophistication of such forms. As more as 21 different measures were compared in the study, all of which used the T-unit as the production unit. The length of productions unit was one of the targets of quantifications when measuring syntactic complexity.

Szmrecsanyi（2004）maintained that the best measure of syntactic complexity was average sentence length, which is definitely an accurate measure. Average sentence length is one of the effective predictor of assessing students' writing, which is confirmed by some researchers. Page (1968), Reid and Findlay (1986) and Slotnick (1972) agreed average sentence length was one predictive aspect of writing performance. Page (1968) mentioned average sentence length correlated with scores at .04 (beta weight=0.09) in the regression model. Other studies have not stated clearly the exact correlation with writing scores, but in their models, it does correlate significantly with scores.

Brown (1994) stated clear difference between writing and speaking on complexity. He claimed that written language have the tendency to use longer clauses and more subordinators.

In conclusion, average sentence length, as well as how sophisticated and varied of the sentence structure can be the predictive of writing performance with regard to complexity.

2.2 Understanding Writing Performance

As is known to all, many researchers use different rubric to assess writing performance. When it mentions to writing performance, scoring rubric and scoring process can not be avoidable. Writing performance always should be presented by scores both in classroom practice and large-scale tests.

2.2.1 Different Criteria of Writing Performance

Writing performance is the evaluation of EFL writing samples by matching the samples with points on rating scale. Kellogg (1994) defined writing performance as "judgments about how well a document communicates or achieves its purpose with its intended audience." Scores can represent writing performance of the essay. Two methods of scoring will be introduced in the following part. Assessing writing performance usually depends on raters making judgments on various kinds of essays on rating scales.

Content, organization, vocabulary, language use and mechanics, these five categories constitute one evaluation scale by Jacobs, Zingkgraf, Wormuth, Hartfiel&Hughey (1981). Later, Isaacson (1988) produced five variables to assess writing: fluency, content, conventions, syntax, and vocabulary. Purves (1985) concluded a few major criteria, content, organization, style, conventions and communication requirements should be summarized. While Block & de Glopper (1992) pointed out that content, organization, style and convention should be applied, because communication requirements is concerned with adequacy of content, which can be included by other criteria.

Some researchers assessed L2 writing focusing content cohesion and quality whereas others such as Robinson (2001) used fluency, accuracy and complexity to assess subjects' performance.

Liang (2005) used language, content and organization as the elementary criteria to assess argumentative samples in the computer-assisted scoring of Chinese EFL learners' essays.

Table 2.1 is the summary of measures of writing quality in previous study.

Table 2.1 Measures of Writing Quality in Previous Study
(Adapted from Liang 2005)

Major Constructs	Sub-constructs		Measures
Language	Fluency		Essay length(in words)
			Fourth root of essay length
			No. of word types
			No. of sentences
	Appropriateness		No. of prepositions
			No. of articles
	Complexity	Lexical	Type-token ratio
			Index of Guiraud
			Vocabulary Frenquency Profile
			No. of common words
			No. of uncommon words
			Average word length
			SD of word length
			No. of nominalizations
		Syntactic	Average sentence length
			% of long sentences
			% of short sentences
			% of simple sentences
			Readability indices
Content	Topic-relevance		SVD similarity measures
Organization	Global		No. of paragraphs
	Local		No. of Pronouns
			No. of connectives

In this study, the writer regards fluency, accuracy and complexity as three dimensions to assess writing performance of subjects.

2.2.2 Writing Rating Methods

There is no consistent standard to assess a piece of writing. In the field of assessing L2 writing, holistic scoring has been adopted widely in abroad and in china, such as TOFEL, TEM4, TEM8, CET4 and CET6. It can be widely used in large-scaled testing. Another important method of scoring, analytic scoring has drawn more attention in recent years. The following part is a brief introduction to these two different scoring.

2.2.2.1 Holistic Scoring Method

Holistic marking is where the scorer "records a single impression of the impact of the performance as a whole" McNamara (2000). Generally speaking, holistic scoring refers to the whole impression of the essay. A single score is assigned to a student's essay on the basis of an overall impression. Holistic scoring can be replaced by global, impressionistic or integrative scoring. Holistic scoring is widely used in the past years.

Zou (2005) considered that holistic scoring was widely used in large-scale writing tests. As it is called, holistic scoring refers to the rater could mark the essays after reading it quickly according to the specific scoring scale or scoring rubric.

A lot of researchers have provided many advantages of this method. Firstly, it is economical to evaluate each essay. Weigle (2002) considered that the essay need to be given only one score based on the total impression after reading only once. Secondly, it is rapid and efficient. It is widely used in large-scaled tests, because this kind of marking can save a great deal of time. Third, White (1984,1985) maintained that the raters focused more attention on the strengths rather than weaknesses of each essay, which was suitable for one to appreciate the essay. White also argued that holistic scoring is more valid than analytic method.

At the same time, there exist some disadvantages of holistic scoring. First, it can not offer more effective feedback to students. Hamp-Lyons (1990) stated that holistic marking was severely limited in that it does not provide a profile of the student's writing ability. Specifically speaking, no detailed information is given about the single mark after rating by holistic scoring, which is not better for students' improvement. The student can not know well why the article had received such score, whether or not because of the well organization or language use. Weigle (2002) considered holistic scoring cannot provide diagnosis for

improving teaching. Second, different raters offer different scores because of preference. Freedman (1979) believed that some raters paid more attention to structure of the whole composition. Grobe (1981) regarded that some preferred word choice. Cohen (1994) cautioned that longer essays often have the trend to receive higher marks. Third, Heaton (1990) pointed out that this kind of marking can be unreliable if marking is done under short time constraints and with inexperienced, untrained teachers. Weigh (1993) admitted holistic scoring faced severe validation.

2.2.2.2 Analytic Scoring Method

Analytic scoring is a type of scoring that demand a separate score for each dimension of criteria. For example, students' writing might be scored including following aspects: content, organization, cohesion, vocabulary, grammar, and mechanics.

Jacobs et al (1981) created the famous analytic scales. Five aspects were considered seriously: content, organization, vocabulary, language use and mechanics. These five aspects weighed differently. Weir (1988) developed a scale for TEEP, focusing on communicative effectiveness and accuracy. Hamp-Lyons (1990) regarded Michigan Writing Assessment Scoring Guide was on three aspects: ideas and arguments, rhetorical features, and language control. Zou (2005) stated that raters gave different aspects of scores according to the scoring rubric in analytic scoring. Although scoring rubric can be used both in holistic scoring and analytic scoring, there are different aspects of ranks relating writing ability and detailed instruction about the ranks. Generally speaking, scoring rubric includes content, organization and language. But there are always something in detail

There are a lot of advantages of analytic scoring. First, different scores of each dimension can be more reliable about the final score. Raters have considered aspects as more as possible through rating. Secondly, each score can reflect different performance in every dimension, showing strong points and weak points clearly, and proving more multi-dimensional effects on participants. For example, the examinees can receive useful information from detailed feedback. They can realize strengths and weaknesses of a piece of writing, which is helpful for them to improve writing quality. At the same time, the researcher can benefit a lot from the final score because more information has been provided from the

final score. Thirdly, Francis (1977) expressed the opinion that new raters can master the standard relatively easily through raters' training. Inexperienced raters can subsequently rates different level with each criteria. Murphy (1982) reported data to confirm this.

The disadvantage of the analytic method is as follows. On the one hand, it lacks of maneuverability in the large-scaled test. Because teachers want to offer more aspects of the same essay, it needs more time than holistic scoring, and then it wastes a lot of labor to some degree. On the other hand, too much attention was drawn from the individual aspect of the composition, which may reduce the overall effect of the writing. Hughes (1989) claimed that a composition score may be very reliable but not valid.

2.2.2.3 Comparison of Holistic Scoring and Analytic Scoring

Holistic or analytic scoring has attracted much attention during the past years. Quite a lot of researchers have conducted empirical study regarding these two different scoring.

The Bachman and Palmer's (1996) framework of test usefulness can be relevant in helping teachers decide which type of test to use. This framework proposed six qualities of test usefulness: Reliability, Construct Validity, Authenticity, Interactiveness, Impact, and Practicality. Bachman and Palmer suggested that test developers develop an appropriate balance among these qualities by setting minimum acceptable standards.

Weigle (2002) commented on the Bachman and Palmer's (1996) framework by showing a comparison of holistic and analytic scales based on the same six qualities of test usefulness as follows:

According to the Table 2.2, reliability and validity are the most important criteria to measure a test.

As for reliability and validity of scoring, there is no definite conclusion about which method is the most reliable one. It is also the controversial question in the field of language testing.

Different research tools have been used in empirical studies about the reliability of scoring methods. As for empirical study, several experts have made research about two different scoring methods. They have got different findings due to different subjects and research tools.

Table 2.2 A Comparison of Holistic and Analytic Scales

(adapted from Weigle 2002)

Quality	Holistic Scales	Analytic Scales
Reliability	lower than analytic, but still acceptable	higher than holistic
Construct Validity	assume that all relevant aspects of writing ability develop at the same rate and can thus be captured in a single score; correlate with superficial aspects such as length and handwriting	more appropriate for L2 writers as different aspects of writing ability develop at different rates
Practicality	relatively fast and easy	time-consuming; expensive
Impact	single score may mask an uneven writing profile and may lead to misleading placements	more scales provide useful diagnostic information for placement and/or nstruction; more useful for rater training
Authenticity	White(1995) argues that reading holistically is a more natural process than reading analytically	Raters may read holistically and adjust analytic scores to match holistic impressions
Interactiveness	n/a	n/a

Prater and Padia (1980), Veal and Hudson (1983), Vacc (1989), Bacha (2001) analyzed the two scoring methods mainly through correlation. All of them came to the conclusion that high correlation coefficient agreed between two scoring method. Anyway, two scoring methods achieve very high correlation coefficients. Both holistic and analytic scoring can provide reliable and valid information.

Cooper (1984) claimed that holistic scoring was of great reliability when the raters were

under severe training. When raters used holistic scoring, inter-rater reliability may be .90

On the contrary, more researchers had provided different results from Cooper. Veal and Hudson (1983) reported the interreliability for raters of .90 by using analytic scoring. For holistic scoring, it ranged from.69 to .76. The main finding is that analytic scoring had highest correlation in assessment. Hamp-Lyons (1991) argued multiple scores tend to improve reliability, which is accepted in the field of second language testing. Stephen Klein et al., (1998) maintained that analytic scoring provides a more objective score than holistic scoring. They reported higher inter-rater reliability using analytic scoring to assess the tasks. Nakamura (2002) investigated two scoring methods respective strengths and weaknesses by using Facets. He stated that the analytic scale is preferable to the holistic scale, because raters function better using analytic scoring method than holistic scoring through the Rasch model adjustment. Akbar Afghari (2004) conducted the research on EFL learners' performance of speaking concerned the effect of the two scoring methods, arriving at the conclusion that different rating scale can offer different results through ANOVA.

To sum, it is acknowledged that there is no consistent agreement with regard to reliability of two scoring method due to different subjects, tasks, research tools, even raters. There is no doubt that holistic scoring and analytic scoring are two effective methods to assess writing performance in EFL learners' writing.

2.3 Summary of Studies on Writing Form and Writing Performance

It is not difficult to arrive at the following summary through related literature concerning the study.

The first part is about writing form. According to Wolfe-Quintero et al. (1998), in L2 writing, a global picture of language development can be manifested by fluency, accuracy, and complexity. Fluency concerns more about the global production of language, while accuracy and complexity are strong related to syntactic aspect of language. Although accuracy and complexity are interrelated, they have different range of fields.

Some findings can be agreed as follows what have mentioned above were most about English native speakers' writing.

First, several researches have made attempts to investigate how the dimensions—fluency,

accuracy and complexity—presented in a piece of essay. Most of them admitted that the three dimensions can reflect one's writing performance, although there is no figure to make a distinction among these three dimensions. At present, there is no clear picture about the dimensions affecting EFL learners' writing. There is much room for experimenting in this field.

Second, the measures of fluency, accuracy and complexity rely on linguistic corpus. As a matter of fact, one important way of assessing writing performance lies on raters. Whether raters can objectively provide fair scores remains a question for further study. In China, raters play a crucial role in classroom teaching as well as in large-scaled tests. One of the important reasons is lack of auto-assessment system. Therefore, it is necessary to test the predictors of writing performance in order to meet demands of current situation.

Third, according to the previous study, there is no respond from EFL about writing form and writing performance. Learners' acquisition can play crucial role in output concerning writing.

In conclusion, fluency, accuracy, and complexity are generally used in assessing both EFL oral practice and EFL written task. Wolfe-Quintero et al. (1998) compared the results of existed research to investigate the best measures indicating second language development. Most researchers had used correlations, multiple regression, ANOVA to provide results for hundreds of measures of fluency, accuracy, and complexity.

The second part is about writing performance. According to what have mentioned in the above, the following conclusion can be reached: previous studies have made some approval of measuring writing performance of certain essay.

Another aspect of assessing writing performance should be noticed from reviewed literature is that the factors affecting scoring in rating. It is easy to find out that text length and average word length have important effects on raters' judgment, while the number of paragraphs, the punctuation, the initial letter of the words, the vocabulary levels of learners may also have a minor role to play in predicting the writing performance.

However, the previous studies have not provided solutions to the following problems:

Firstly, fluency, accuracy and complexity has been widely utilized by researchers to

assess oral discourse and written course, although a few researchers had involved other aspects such as content, organization and mechanics in the rating scale. There is no agreement in rating.

Secondly, how can these three dimensions—fluency, accuracy, complexity—predict writing performance remains a question. To what extent can these three dimensions predict more about the writing is not seen clearly. During the previous study, a few researchers reported correlation coefficients in regressional model about only one or two dimensioned. However, in study, correlation coefficients among three dimensions are not known clearly.

Thirdly, nearly all the studies probe into finding correlations between essay features and writing performance by native speakers of English. More studies need to be conducted about EFL writers' essays, the existing measures and some new measures should be carried out in assessing EFL writing performance.

2.4 Empirical Studies on Structural Equation Modeling

SEM approach, unlike multiple regression analysis, is desired for simultaneous analysis of both observed and latent variables. It is able to assess and account for measurement errors. In a word, it is a powerful method.

SEM is one way of analyzing variables, having more advantages than of Confirmatory factor analysis and Path analysis. Confirmatory factor analysis can analyze and test the cause and effect between observed variables and construct. Path analysis can analyze and test the relationship among observed variables. While SEM can probe into the inter and intra relationship of many observed variables.

SEM has been widely used in applied linguistic recently and it has a prospect future. Especially in individual differences in SLA (Gardner, Lalonde & Pierson 1983; Gardner et al. 1987; Gardner 1988; Clement& Kruidenier 1985; Ely 1986); individual differences (Csizer & Dörnyei, 2005; Lee, 2005; Sasaki,1993; Gelderen, Schoonen, et al, 2003;Woodrow, 2006; Yashima, 2002; Yashima,Zenuk-Nishide & Shimizu, 2004), language testing (Purpura, 1998; Schoonen, 2005).

Schoonen (2005) conducted the study concerning testing by research tools SEM. In Schoonen's study, it was extremely difficult to assess writing ability. Its outcome can be

affected by different dimensions of the criteria. The score of a piece of writing can be decided by the following facets: the topic of the writing, the way of scoring, the raters selecting, the scoring rubric, including content, language, organization, and so on. Two scoring methods have been used for raters and two different traits i.e. content and organization, and language use have caught attention by the raters. In the study, SEM was used to estimate the variance in the score. The conclusion was as follows: holistic scores were more generalizable than analytic scores collected with scoring guides.

Purpura (1996) examined the relationship between the application of metacongnition and language testing for second language. It was one of the important studies in language testing.

Experts in China have adopted SEM as effective way to do research, especially in SLA and language testing.

Wang Lifei (2004) established a model to interpret the effects of L1 literacy variables on L2 writing ability through LISREL. Wu &Liu, (2004) examined the effects of metacongnition in EFL writing, investigating the relationships between writing metacognition and writing performance by SEM.

Wang Lifei (2003, 2007) pointed out that the importance of SEM to applied linguistic research, and introduced three different function of SEM through the software LISREL.

Han (2006) stated the application of SEM in language testing. He introduced SEM as a way of examining language testing. He discussed the advantages of SEM over other techniques and the prospect future of SEM in language testing.

Wen (2007) measured the validity of the construct "content" by using SEM. She collected 120 argumentative essays. She introduced four observable variables: relevance, explicitness, coherence and sufficiency. She assumed these four observable variables can indicate content. SEM was used to establish the model for explaining predictability of writing performance about the four variables. The ultimate purpose of the study was to provide a better analytical scoring guide for rating the content quality.

In conclusion, SEM, as one advanced analysis tools, can provide more scientific research in language testing. SEM can offer a number of advantages over previous multivariate procedures (Wang Lifei, 2004).

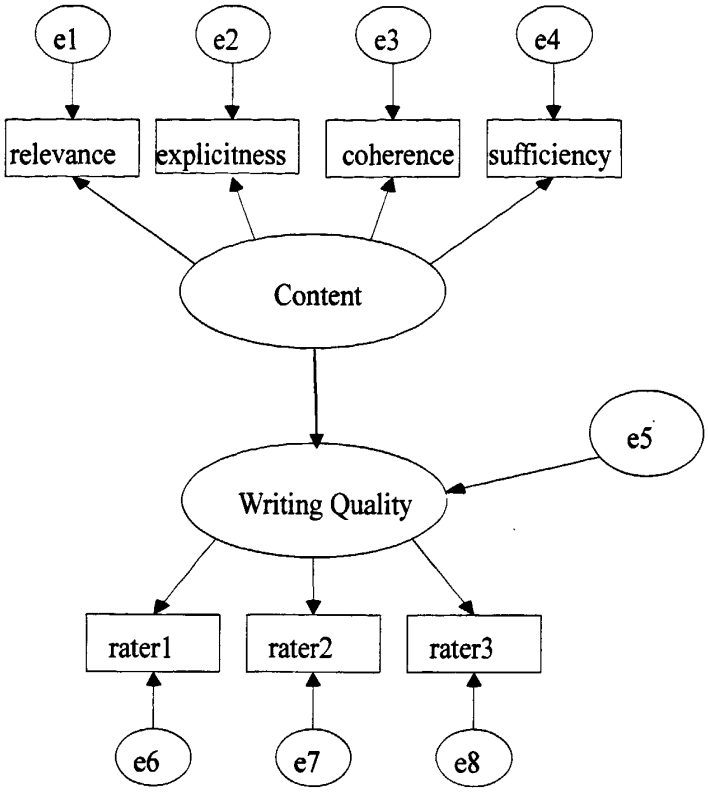

Figure 2. 1 Measuring the Validity of the Construct "Essay Content "by Using Structure Equation Modeling
(Adapted from Wen 2007)

In order to probe into different contributions of writing form to writing performance, priority is given to SEM in this study.

Chapter Three Methodology

In this chapter, objectives of the present study are presented in 3.1. Then in 3.2 subjects are introduced in detail. In 3.3 essay writing and interview, as two main instruments of the study, are provided thoroughly. 3.4 is a brief introduction of data collection. At last, in 3.5 rating scale and rating procedure are given elaborately.

3.1 Research Questions

The present study aims at probing into the relationship between form and writing performance, and even different predictability of three dimensions to writing performance. At the same time, SEM has been constructed to validate the findings.

The study would try to answer the following questions:

1. In L2 writing, what is the relationship between form and writing performance?

2. As for the L2 learners, what are the contributions of three different dimensions to the writing performance? Which predict more about writing performance, fluency, accuracy or complexity?

3. As for the high proficiency learners and low proficiency learners, what is the relationship and obvious tendency of the different dimensions to the writing performance? How the understanding of the different dimensions to influence the writing performance?

In the quantitative part, the first two questions could be solved by data analysis, while in the qualitative part, interview was employed in the study to give more insights in the third question, namely, according to interviewees' response to interview questions, is there any difference of the tendency between form and writing performance with the improvement of students' writing level?

3.2 Subjects

There are 243 students who have been required to write this article in 50 minutes during the final examination. All of the students majoring in English come from Grade Three of different universities in Hebei Province. They are about twenty-one years old. They are selected because the students had finished learning writing course in the previous semester,

and most of them had passed TEM4. They were expected to represent the advanced level of EFL majors.

Table 3.1 Demographic Information of Subjects

Category	Description				
University	3 universities in Hebei				
Major	English				
Grade	Three				
Average age	21 years old (ranging from 19 to 22)				
Literacy level	1.3rd year of college English learning (about 9 years of English learning) 2.Most passed the TEM4				
Sample	Total sample	246			
	Missing sample	3			
	Valid sample	243	A University (comprehensive)	Two classes	75
			B University (science and technology)	Two classes	69
			C University (economy)	Three classes	99

All the students are native speakers of Chinese. Most of them had begun to learn English from primary school. They take the composition seriously and they are careful since they should finish the article in the examination. In other words, they are eager to show the best performance in the final examination, therefore they are cooperative to finish the writing task to a great degree. Although all of the students participate in the final examination, three of the compositions are missing samples because of few sentences in the composition.

3.3 Instruments

3.3.1 Essay Writing

In the quantitative part of the study, one argumentative English essay should be written in the limited time by the students who major in English.

The argumentative writing is used as the topic for the students due to the following reasons. Firstly, as one of the types of composition, argumentative writing can test the ability of students in more multiple aspects. Argumentative writing can express one's opinion concerning one certain topic. For example, argumentative writing can display the ability of analyzing and solving problems, the ability of logical reasoning and the organization of ideas, as well as the ability of language use. Secondly, argumentative writing can test the thinking progress as well as the application of language. Argumentative writing provides more requests about cognitive and thinking. Thirdly, argumentation can take the cognitive competence into consideration. In addition, argumentative writing can probe more into the critical thinking of each student.

The original direction is as follows:

Directions: It has been reported that more college graduates tend to choose to work as village officials in the countryside. To their choice people's attitudes differ greatly.

What is your opinion on the topic? Please write a composition of no less than 350 words within 50 minutes on the following topic:

College Graduates Work as Village Officials

The topic is chosen for the following two reasons. Firstly, as is known to all, village official has been a hot topic in recent years. As the college students, especially the students who are in Grade three, they are concerned with the problem of employment. Secondly, for this topic, content would not become the problem for subjects. As the junior, they must have known something related to this topic from TV, newspapers, magazines and websites. They have learned a lot from the media even the experience of their friends or relatives who had been chosen as a village official after graduation. Therefore, they must have had a lot to write about the article. In a word, this topic is popular and practical for each student.

In this empirical study, SPSS (16.0) play an important role. All of the data collected can

be analyzed according to SPSS (16.0). These raw data were put into SPSS (16.0) to check the reliability of rating, the correlation of content, form even the performance of writing and so on. AMOS (7.0), this software, was used to establish the SEM to reveal the relationship among different variable clearly. It is easy to read the raw data into the model. According to the model, it is easy to read and analyze the results through the raw data.

3.3.2 Interview

In the qualitative part of the study, the purpose is to illustrate and confirm the findings in the quantitative part and gain deeper insights into the relationships of fluency, accuracy, and complexity in the compositions.

The analysis involves interview. Interview is one approach of qualitative sources. Interview is conducted in order to know something truthful from the students' perspective. The interview can pursue in-depth information around the topic. In this study, the research have adopted in-depth interview because it involves conducting intensive individual interviews with a small number of respondents to explore their perspectives on a particular ides, program, or situation. (Carolyn, 2006)

The researcher have selected four subjects and interviewed them one by one. The selected subjects were two high proficiency learners and two low proficiency learners according to the writing performance. The researcher told the subjects that all of the conversations had been recorded in the interview in order to conduct the study. The researcher also told the subjects not to be tense and nervous. The researcher should be a good listener in the interview. During the interview, the researcher had asked the subjects a series of questions that prepared well in advance in Chinese. The questions concerned the process of writing; whether they show emphasis on form; the selection of words or patterns of sentences; whether they paid more attention to choosing words and sentences. All the selected subjects can answer the questions as elaborately as possible.

3.4 Data Collection

The data were collected in different time because the subjects come from different universities. The data collection was first undertaken in Hebei University of Economics and Business and then in Hebei University and Hebei University of Technology. The process of

collecting had been controlled seriously during the final examination. Anyway, they were cooperative to have such a test in order to get high marks in the final examination.

3.5 Rating

3.5.1 Holistic and Analytic Scoring Scales

All of the compositions need to be scored twice, one with holistic methods, another with analytic method. The researcher intended to use the results of holistic rating as the writing performance of each composition. On the contrary, analytic rating used to provide more feedback of the compositions that the students have written, which can illustrate well about the sub-components: content and form. The former includes relevance, explicitness, coherence, sufficiency and the latter includes fluency, accuracy, and complexity.

On the one hand, as for holistic scoring scale, the researcher decided to use the standard of evaluating TEM4, because the subjects come from Grade Three, and most of them have participated in the examination of TEM4. To some degree, it is fair to every student by using the criteria of assessment. What's more, it is easy for the raters to assess the compositions with relevance to this type of rating. It is acknowledged that the teachers had adopted this scoring criteria to identify the students' writing proficiency level during teaching. During the examination, it is very common to be used to evaluate students' writing performance. Anyway, holistic scoring is widely used both in classroom teaching and in tests. One thing should be pointed out, when rating, the score is ranging from 0 to 100, which is of little difference from that of TEM4 because writing occupies 15 points in TEM4.

On the other hand, as for analytic scoring scale, after careful study of some widely used criteria, the researcher chose to employ the criteria which had been proved valid. (Wolfe-Quintero et al., 1998; Liang, 2005 Wen, 2007). As for criteria, it can be included by content and form according to its features. There are four dimensions concerning content and three dimensions concerning form. Admittedly, form was the focus of the study whereas content used to help analyze the result.

First, content is composed of relevance, explicitness, coherence, and sufficiency. Relevance to the topic is critical. If the student can not understand the topic properly, the composition who wrote will be far away from the topic. Explicitness refers more to the topic

sentence. The raters focus on whether the topic sentence is clearly expressed or not. Coherence shows concern to the relationships between and inter course, including the relationship between sentences, paragraphs even the whole passage. Sufficiency implies whether there is enough grounds of argument supporting the argument thesis.

Form can be made up of three dimensions—fluency, accuracy and complexity. These three indicators have been proved as effective measures to assess L2 writing by western researchers Wolfe-Quintero et al. (1998). In the study, fluency focuses on the number of total words and readability. Accuracy refers to total error, including misuse of grammar, misspelling and so on. Complexity includes lexical complexity and syntactic complexity, referring to advanced words and complex sentence pattern.

Specifically, fluency can be assessed based on the length of the article and fluency of expression. There is the exact total word number on the composition from Microsoft Word. Accuracy is based on grammatical accuracy, and error use of word, paragraphing, spelling and even punctuation. Accuracy is concerned with proper expression, including the structure of the sentence, the usage of subject predict, tense, collocation, Chinglish and paragraphing. Spelling, punctuation, plural forms and so on are also included in this aspect. Complexity is based on sophistication of vocabulary range, effective word choice and usage, complex patterns of sentences, such as inverted sentence, attributed clause, adverbial clause and so forth.

After comparing with other standards, the researcher believed that these three dimensions can reflect well of the composition. Each dimension weights equal, ranging from 0 to 50. Although the large gap of each level may bring difficulties in understanding and scoring, the lag gap can distinguish high proficiency learners from low proficiency learners well. Another advantage was that the detailed data can provide more feedback to the researcher and the subjects. The raters can give each score according to the compositions of the subjects. The average of the scores is regarded as the final score of each subject.

3.5.2 Rating Procedure

It is crucial to mark the composition. Before scoring, the author input all the articles into the computer. The author didn't correct the mistakes, for example, spelling, punctuation,

paragraphing and grammar when the articles were typed into the computer. Meanwhile, the number of the words of each article was calculated by the Microsoft Word to get the length of every composition and in order to let the raters know the exact words of the document. When typing into the computer, the name, the class and the university had been omitted in order to avoid some kinds of possible effect on raters. Then all of the compositions were divided into eight groups randomly, each having thirty samples. Therefore, each composition was assigned a number of group and a number of inter group, which was helpful to find the composition when doing analysis. At last, all of the papers should be copied to each rater. What have mentioned above can create a natural environment of rating. At the same time, it can reduce the negative effects on rating, such as the impact of handwriting, the different level of the universities.

The researcher chose six experienced teachers to evaluate the articles. The researcher was one of the raters. All of the raters were experienced college English teachers for several years, ranging from seven years to sixteen years. All of the teachers are professors and got master degree except for the author. They had rich experience of teaching and they were qualified teachers and popular among the students.

All the task of rating began from Late January in 2010. There was no teaching task in the winter vacation, so each rater had sufficient time and energy to ensure the speed and quality of rating. At last, all the results were put into SPSS (16.0) to do analysis statistically.

Three of them should give the holistic score, while others should give the analytic score, that is to say, three of them should give only one score according to the standard which is mentioned above. The other three raters should give a score regarding seven dimensions on the basis of Wen (2007) and Liang (2005) formulated. The three scores are averaged as the final score of the two parts of the compositions.

During scoring, two points should be paid more attention.

First, it is very important to train the raters before scoring, especially the analytic scoring. The score is affected by personal beliefs or preferences because rating is done by human being. Training before rating is to decrease the biases or differences caused by different raters. As for holistic scoring, it is relatively easier to master. The training process is not complicated.

Three of the raters learn the criteria together. They gave the scores for three compositions respectively. The inter-scorer reliability was calculated and was found high ($r=0.78$). As for analytic scoring, in order to ensure validity of the evaluation, it's necessary for the raters to discuss, revise and understand the standard before scoring. At first, the raters had been required to the meeting room and discussed the standard together to arrive at a great agreement in terms of understanding the specific criteria, such as how to understand this topic about relevance to the thesis of the topic, clearness of the statements. Then they had assessed three compositions individually. After rating, they reported the score one by one. The inter-scorer reliability was calculated and was found very high ($r=0.81$). In a word, the training aims to let the raters adopt the new criteria of scoring, avoiding their own scoring experience in assessing the articles, especially the analytic scoring, which is not widely used in daily teaching.

Second, after training, the raters should assess the article respectively. When they meet the difficulties, namely, it is hard to decide the score, they can regard the composition as particular one, discussing with other raters. If they disapprove of the final score, they should discuss in order to get the consistent mark. If they can not achieve the agreement after discussion, they can keep the original score. When the score is highly different from others, this article should be read again so as to get the fair result.

In conclusion, rating is one of the key points in the study. It is glad to find that all of the raters are very responsible for the task. All of the raters have committed themselves to this task during rating. Although the raters perhaps may be influenced by some unavoidable factors, for example, time, unpleasant feeling or different views regarding the same composition, the reliability can be ensured to a great degree.

Chapter Four Results and Discussion

In this chapter, the results of the data and discussion to each question have been analyzed in detail. The present study is mainly a quantitative study. SPSS (16.0) is used to conduct data analysis, including score reliability and correlation analysis. Correlation analysis are employed to compare the results of holistic scoring and analytic scoring. AMOS (7.0) is used to construct SEM concerning the topic. In discussion part, three research questions have been put forward in Chapter 3 discussed in detail. In this part, interview is adopted to compare and give deep insights to different performance between high proficiency learners and low proficiency learners.

4.1 Data Presentation

This section reports the results of the data analysis by SPSS (16.0). All of the data were put into the computer before using SPSS (16.0) to analyze.

4.1.1 Score Reliability

In this study, a few score reliability were reported. Reliability report was to see whether the raters have provided reliable and effective scores for the essays in the study.

Intra-rater reliability and inter-rater reliability constitutes score reliability.

Intra-rater reliability is provided to test self-consistency, while inter-rater reliability is produced to express agreement between raters, just as Hyland (2003) argued agreement on the rating of the same learner performance is intra-rater reliability. Objectively, each rater should perform consistently on different occasions during the assessing procedure.

Table 4.1 Correlations among Raters in the Holistic Scoring and Analytic Scoring

	Crobach's Alpha	N of Raters	N of Samples
Holistic scoring	.818	3	243
Analytic scoring	.802	3	243

Table 4.1 is about Crobach's alpha coefficients, which is another mean to compute consistency of inter-rater reliability. Crobach's alpha is 0.8 which is positive evidence of

agreement. Table 4.1 shows respective correlation coefficients among raters in holistic scoring and analytic scoring. If the correlation coefficient ranges from 0.70 to 0.90, it is regarded as high correlation. In the study, three raters in different approach of scoring have achieved high inter-rater correlation, 0.818 and 0.802 respectively. Although there is a little difference between two scoring methods, both of them lies to the range of 0.70-0.90, implying very high correlation among raters.

4.1.2 Correlation Analysis

Weigle (2002) mentions a correlation coefficient between scores assigned by different raters is regarded as one method of analyzing score reliability or consistency. A correlation coefficient close to 0 implies there is little of no relationship between the scores, whereas it close to 1 indicates a strong relationship between the scores.

At the same time, correlation analysis was used to report the relationships of different dimensions, including correlations among raters in the holistic and analytic scoring respectively, the correlation between holistic scoring and analytic scoring, as well as means and deviations of different score. Correlation analysis is conducted in order to reveal the reliability of rating and whether there is relationship between form and writing performance.

Rater1, Rater2 and Rater3 provide holistic scoring, while Rater4, Rater5 and Rater6 offer analytic scoring.

Table 4.2 Correlations among Raters in the Holistic Scorings

		Rater1	Rater2	Rater3
Rater1	Pearson Correlation	1.000	.860**	.825**
	Sig. (2-tailed)		.000	.000
	N	243.000	243	243
Rater2	Pearson Correlation	.860**	1.000	.755**
	Sig. (2-tailed)	.000		.000
	N	243	243.000	243
Rater3	Pearson Correlation	.825**	.755**	1.000
	Sig. (2-tailed)	.000	.000	
	N	243	243	243.000
**. Correlation is significant at the 0.01 level (2-tailed).				

Table 4.3 Mean and Standard Deviation of Holistic Scoring

	N	Minimum	Maximum	Mean	Std. Deviation
Rater1	243	25.00	94.00	64.8107	12.47233
Rater2	243	15.00	92.50	63.6975	16.21008
Rater3	243	32.00	90.50	65.7058	11.95104
Valid N (listwise)	243				

Table 4.2 reports correlation among raters with regard to different scoring by Pearson Correlation. Pearson Correlation is the most common measure of approvement among multiple raters. Page (1995) says ETS rating experts typically correlate at 0.5-0.6. That is to say, it is acceptable for the average correlation to 0.5-0.6. In this study, Correlation among raters is 0.860, 0.825 and 0.755. The average correlations with three raters by holistic scoring are r=0.813. Noticeably, the results of writing performance are highly reliable.

As is shown in table 4.3, there is a little difference among three raters, indicating that raters have understood the criteria deeply and similarly with much agreement. Therefore, they can perform well during the rating procedure. To some degree, three raters have mastered the rating scale and provided fair score.

Table 4.4 Correlations among Raters on Fluency in the Analytic Scoring

		Rater4	Rater5	Rater6
Rater4	Pearson Correlation	1.000	.730**	.756**
	Sig. (2-tailed)		.000	.000
	N	243.000	243	243
Rater5	Pearson Correlation	.730**	1.000	.779**
	Sig. (2-tailed)	.000		.000
	N	243	243.000	243
Rater6	Pearson Correlation	.756**	.779**	1.000
	Sig. (2-tailed)	.000	.000	
	N	243	243	243.000

**. Correlation is significant at the 0.01 level (2-tailed).

Table 4.5 Mean and Standard Deviation of Fluency Scores

	N	Minimum	Maximum	Mean	Std. Deviation
Rater4	243	11	47	35.93	6.465
Rater5	243	17	48	35.45	6.479
Rater6	243	14	48	34.66	7.490
Valid N (listwise)	243				

As can be seen clearly, Table 4.4, and Table 4.5 illustrated correlation among raters of fluency scores in terms of analytic scoring by Pearson Correlation. As for fluency, three raters correlate at r=0.730, r=0.756, and r=0.779 respectively, average correlation is 0.755, a little higher than 0.5-06, which is positive results. The standard deviations of the three sets of scores are also so close that indicate three raters have used similar rating scale.

Table 4.6 Correlations among Raters on Complexity in the Analytic Scoring

		Rater4	Rater5	Rater6
Rater4	Pearson Correlation	1.000	.512**	.583**
	Sig. (2-tailed)		.000	.000
	N	243.000	243	243
Rater5	Pearson Correlation	.512**	1.000	.564**
	Sig. (2-tailed)	.000		.000
	N	243	243.000	243
Rater6	Pearson Correlation	.583**	.564**	1.000
	Sig. (2-tailed)	.000	.000	
	N	243	243	243.000
**. Correlation is significant at the 0.01 level (2-tailed).				

Table 4.7 Mean and Standard Deviation of Complexity Scores

	N	Minimum	Maximum	Mean	Std. Deviation
Rater4	243	15	47	24.19	7.121
Rater5	243	15	43	22.32	6.025
Rater6	243	15	43	22.78	6.065
Valid N (listwise)	243				

Table 4.7 shows that the correlation among raters on complexity is 0.512, 0.583 and 0.564, and average correlation is 0.553, a bit lower than correlation among raters on fluency, however it is also acceptable for the study. All in all, Pearson Correlation implies that the scores that the raters assigned are at significant level. As for the standard deviations, the three sets of scores are also close.

Table 4.8 Correlations among Raters on Accuracy in the Analytic Scoring

		Rater4	Rater5	Rater6
Rater4	Pearson Correlation	1.000	.599**	.711**
	Sig. (2-tailed)		.000	.000
	N	243.000	243	243
Rater5	Pearson Correlation	.599**	1.000	.745**
	Sig. (2-tailed)	.000		.000
	N	243	243.000	243
Rater6	Pearson Correlation	.711**	.745**	1.000
	Sig. (2-tailed)	.000	.000	
	N	243	243	243.000
**. Correlation is significant at the 0.01 level (2-tailed).				

Table 4.9 Mean and Standard Deviation of Accuracy Scores

	N	Minimum	Maximum	Mean	Std. Deviation
Rater4	243	5	50	30.93	10.219
Rater5	243	5	49	23.19	10.308
Rater6	243	4	49	25.63	11.178
Valid N (listwise)	243				

Table 4.8 shows that correlation on accuracy scores range from 0.599, 0.711 to 0.745. And average correlation is 0.685. Table 4.9 shows that there is a little difference among three raters, suggesting that the raters the raters have followed the criteria well because of the similar scores they have assigned according to close ranges of mean and the standard deviations of scores.

As can be seen clearly in the above tables, the statistics show that the holistic scoring and analytic scoring have produced positive results.

Both holistic and analytic scoring could be used to apply in this study. Although there is controversy about reliability and validity of holistic and analytic scoring in the related literature review, the results are rather satisfactory in the study.

4.1.3 SEM Analysis

SEM can be utilized to analyze the relationship among variables that have relations. In this study, SEM can be applied to test the relationship between form and writing performance.

4.1.3.1 Constructing SEM

Fluency, accuracy, and complexity are three observable variables in the model. The hypothesis is as follows: if fluency, accuracy, complexity can reflect writing form, at the same time, writing form can interpret writing performance, then writing form can predict writing performance. Meanwhile, fluency, accuracy and complexity can be three predictors to writing form.

Rectangle stands for observed variable; ellipse stands for latent variable; circle is for error variables or residual. One-way arrow stands for the relationship between latent variables, implying latent variables can be reflected by observed variables. One-way arrow can also

refer to the relationship among latent variables. Starting point of the arrow is the independent variable and terminal point is dependent variable. Coefficient in each path indicates to what extent one latent variable affect another, namely predictability or explanatory power.

This model is composed of two measurement models. One is composed of latent variable: form and three observed variables—fluency, accuracy and complexity as well as three error variables. Another one regards writing performance as latent variable, different scoring of raters as observed variables, three error variables and one residual.

The preliminary model is as follows:

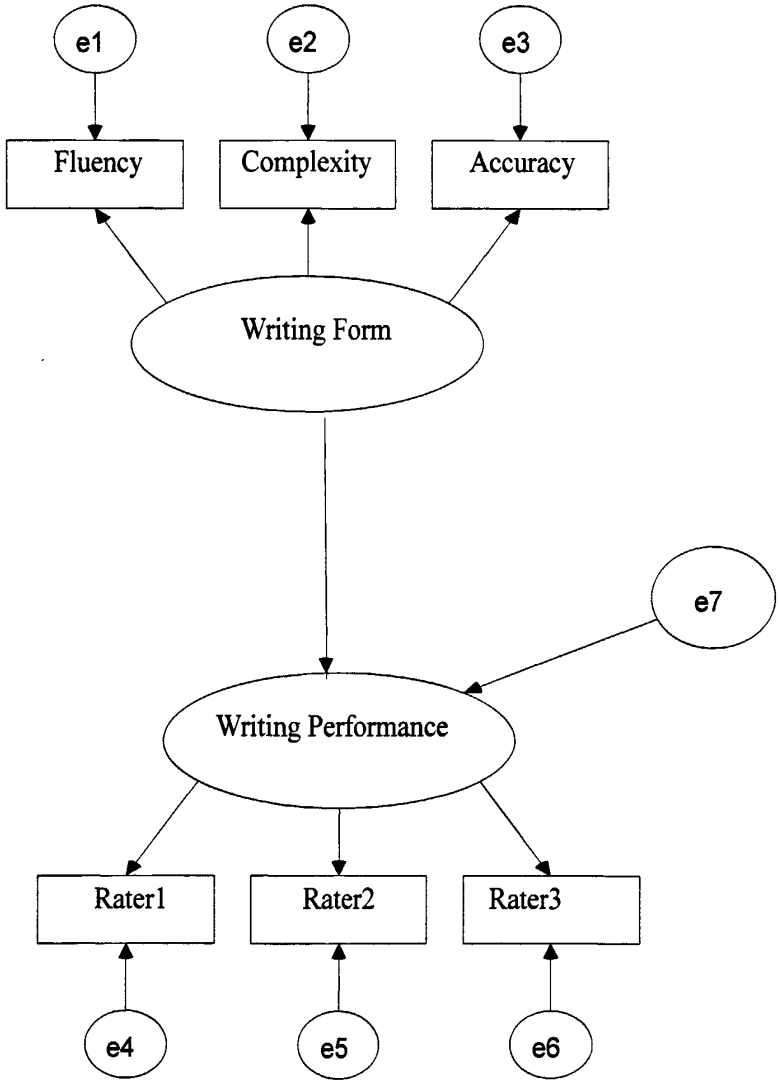

Figure 4.1 Preliminary Model

4.1.3.2 Adjusting Models in the Study

After putting the original data into the model and doing operation by AMOS (7.0), some indices should be modified according to the suggestions given by the software.

Figure 4.2 the refined model 1 indicates that according to the suggestions from AMOS (7.0), the writer used double-way arrow to show covariance between two variables, e2 and e3. After operation of AMOS (7.0), this model is not the perfect one because some indices can not fit the model well. The specific data are presented in Table 4.10.

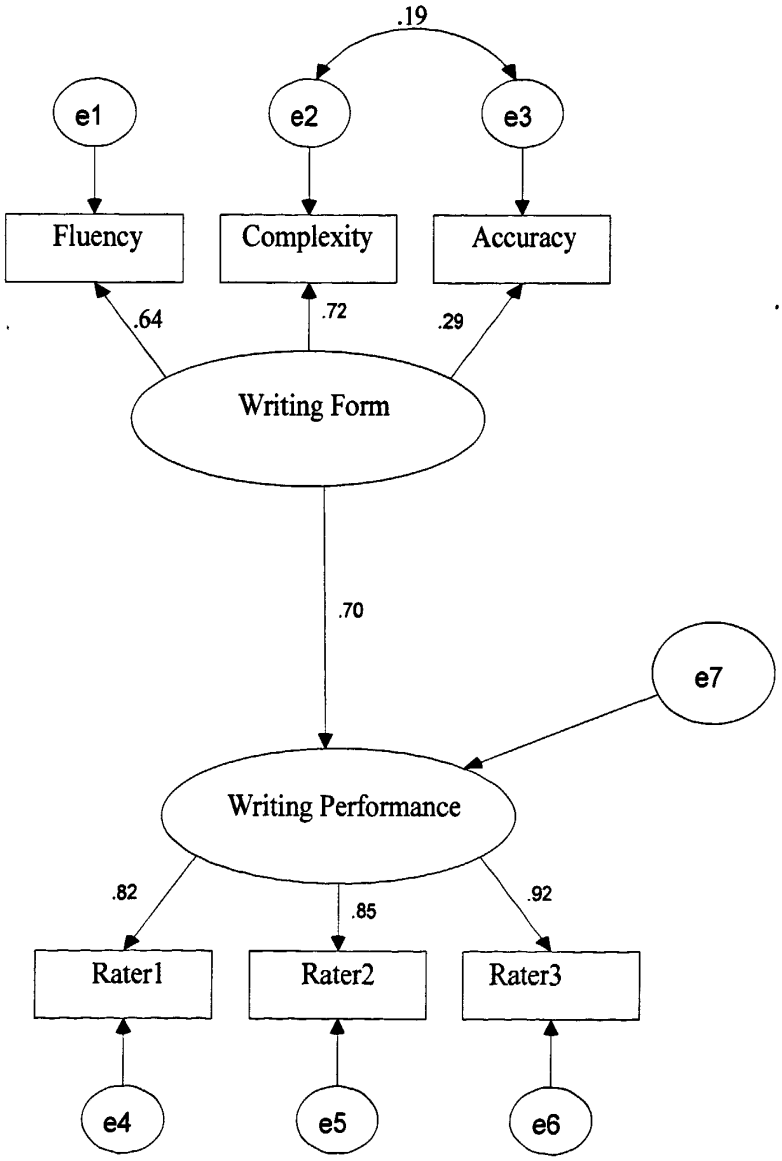

Figure 4.2 Refined Model 1

According to the suggestions of the software, the writer put the double-way arrow between e3 and e5 to show covariance between these two variables. There are double-way arrow between e2 and e3, e3 and e5. At last, the indices achieved statistics requirements after operating AMOS (7.0).

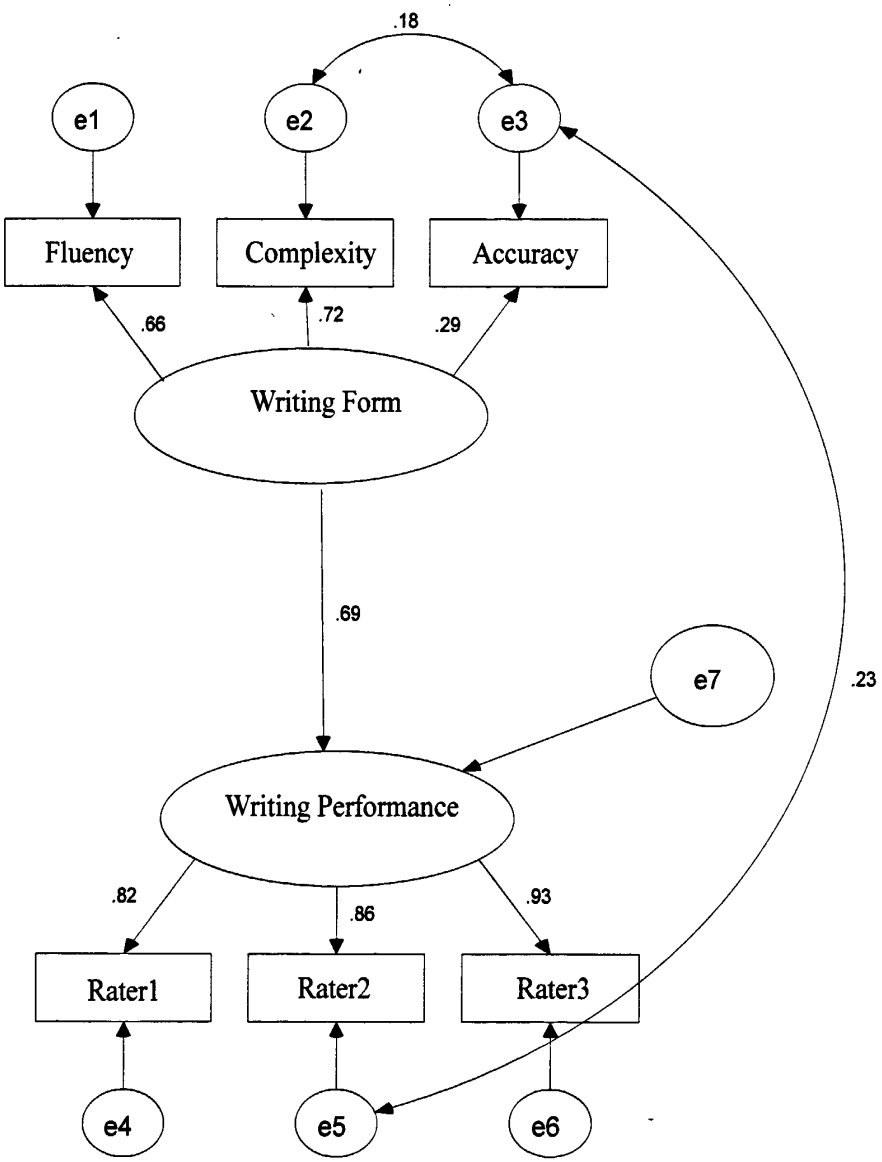

Figure 4.3 Refined Model 2

Table 4.10 Model Fit Summary

	Chi-Square	DF	P	GFI	AGFI	CFI	RMSEA
Preliminary Model	29.655	8	.000	.958	.891	.969	.104
Refined Model 1	28.623	7	.068	.963	.921	.975	.093
Refined Model 2	11.494	6	.074	.985	.946	.992	.061

Table 4.10 shows fitting effect between preliminary model and refined model. When Chi-Square to degree of freedom<2, $P>.05$, GFI, AGFI, CFI$\geq .90$, RMSEA$\leq .08$, the model can be accepted. In the preliminary model, GFI is .958, CFI is .969, both of which meet the requirements because the data is above .90, whereas Chi-square to degrees of freedom>2. and p is 0.00, which is less than 0.05, AGFI is $.891<.90$ and RMSEA is $0.104>.08$, these data had no statistical significance. The statistic showed in refined model 1 is not satisfactory. Because Chi-Square to degree of freedom>2 and RMSEA is $.093>.08$, both of which can not meet the needs either. According to the suggestions from AMOS (7.0), the writer used two double-way arrows to show covariance between two variables, between e2 and e3, e3 and e5. After refining the model again, the refined model 2 has met with requirements of statistics: Chi-square is 11.494, DF is 6, P is .074, GFI is .985, AGFI is .946, CFI is .992, RMSEA is .061, all of data meet the statistical requirements.

4.1.3.3 Interpreting SEM

According to Figure 4.3, the path coefficient of fluency, complexity and accuracy of writing performance are .66, .72, .29. The differences refer to their respective contributions to the writing form, anyway, all of the data is effective index, which can reflect the performance of writing to some extent. Among these, complexity can contribute more than the other two variables because of higher figure to writing form, whereas accuracy is the least contributor to writing form due to the least figure. In practice, raters paid more attention to vocabulary and diversity of sentences instead of errors during scoring, which leads to the above result.

With regard to rater1, rater2, and rater3, the scoring by holistic scoring method, the

respective path coefficient are .82 .86. 93. There are different contributions for the raters in the study. Rater6 has contributed most to the final results.

Although there are differences, all of the data can be of statistics meaning, because factor loading is above 0.70, which implies they can reflect writing performance appropriately. The results can be interpreted as follows: the highest number indicate the highest contribution to writing performance whereas the lowest number can reflect least about writing performance. In fact, there is no same rating in practice because of various kinds of factors affecting rating.

Writing form to writing performance is .69, explaining 47.61% variances of writing performance directly. Wen (2007) found that writing content can interpret 56% variances of writing quality. The results in this study are similar to that of Wen (2007). That is to say, the content including relevance, explicitness, coherence, and sufficiency, these criteria creating by Wen (2007) can reflect more about writing quality. There is a little difference in this study because of the number of subjects and different level of subjects and even different task to subjects. As Robinson (2001) put forward the view that task difficulty and task complexity can affect students' writing performance.

The finding in SEM can prove the following point that fluency and complexity may be more associated with language development than accuracy is. (Wolfe-Quintero et al., 1998).

In a word, the finding in SEM can provide more evidence to teachers and students in L2 writing. At the same time, the findings are helpful for us to improve the way of assessing L2 writing.

4.2 Discussions

In L2 writing, form can be positively related to writing performance. The findings proved the previous studies in some aspects. Wolfe-Quintero et al. (1998) maintained that fluency, accuracy and complexity are the most scientific method to assess writing performance. In this study, these dimensions are really truthful to assess writing performance.

4.2.1 Discussion on Research Question 1

The first question is as follows: in L2 writing, what is the relationship between form and L2 writing scores?

In L2 writing assessment, there is no consistent rubric to perform. In China, the rubric

employed in large scaled test varies from each other.

The method to assess writing in the entrance examination for college, divide into five grades. For the raters, they should classify the article into one certain grades according to content and language after viewing at first glance. The number of words per article is also considered one aspect to assess it. The number of lexical and syntactic sentence should be concerned and accuracy is important. Spelling and punctuation are important aspects to assess accuracy. As shown clearly, it focused more about language form, including fluency, accuracy and complexity.

CET4 and CET6 also use holistic scoring. From relevance to content, clear statement and structure knit well and language standardized. TEM8 also adopt holistic scoring, which attempts to assess from content, language and structure. Content involves clearness and sufficient of the argument; language involves accurate of sentence and vocabulary and even register; structure refers to logical of whole composition and correct use of cohesion.

Wen (2007) points out that existed rubric focused more about language form than content. She rebuilt the rubric to assess L2 writing form encompasses vocabulary, sentence patter, grammar and mechanics. Vocabulary refers to the appropriateness of collation; sentence pattern refers to complexity and diversity of sentence; grammar refers to accurate use of grammar rules and mechanics pay attention to punctuation, spelling and capitalization. Content include the following aspects: relevance, explicitness, coherence, and sufficiency. Wen (2007) had built SEM to measure the validity of the construct essay content. She found that content can predict 56% variance of writing performance.

There is no doubt content and form can be effective factors to decide the performance of writing. Form can also reflect the writing performance, especially for L2 writing. Generally speaking, fluency, accuracy, and complexity can encompass different aspects of a piece of writing. There is close relationship between form and writing performance. Just as data shown, the above three dimensions have positive correlation with writing performance.

Admittedly, the findings in this study had a little difference with that of Wen's (2007) study. One important possible reason is that subjects' English level can determine the writing performance they had presented in the study.

4.2.2 Discussion on Research Question 2

Question 2 is concerning the following one: as for the second language learners, what are the contributions of three different dimensions to the writing performance? Which predict more about writing performance, fluency, accuracy or complexity?

The three dimensions can predict writing performance, but each has its own contributions to writing performance.

Previous studies had reported coefficient to indicate writing performance according to correlation analysis and ANOVA. In the study, SEM is employed to make clear about the predictability scientifically. SEM can examine the variances, which have more advantages than ANOVA. As to this study, SEM is more appropriate and effective than ANOVA.

Firstly, complexity, one of the important predictor, is placed more emphasis on lexical and syntactic richness during scoring by human raters. It is easy to make the clear point that raters give more emphasis to a wide range of lexical richness and complicated sentence pattern during scoring. More measures and the importance of complexity have been reported in the literature review part.

The findings are similar to related studies about complexity. In the present study, complexity contributes more than other two dimensions, indicating that complexity is closely related to the score of composition. In previous studies, some researchers reported the results of their findings about complexity.

McCurry and McCurry (1992) reported the results the number of word types is significant correlated essay scores for native speakers. Kaplan et al.'s (1998) founded essay length and 4^{th} root of essay length have significantly correlation with essay score, referring holistic score. Liang (2005) confirmed the findings. Liang (2005) put forward the view that for L2 writers, syntactic complexity is considered one of the most important aspects to distinguish good writing from poor writing.

Secondly, with regard to fluency, different results have been reported by researchers.

Wolfe-Quintero et al. (1998) employed several objective measures to evaluate fluency, which has quite a high correlation with writing proficiency at .65 or above. Homburg (1984) argued that length measures accounted for 84% of the variation. Length measures refers to the

number of words, the number of T-units, the number of sentences, the number of T-units and words per sentence, and the number of words per T-unit.

In this study, fluency has positive relation with writing performance, which is set at .66, much closer with the findings in Wolfe-Quintero. The possible reason is that although different methods have been employed, they used objective indices while in this study, human raters scoring could be reliable to examine, subjective matters can not be totally avoided. Fluency can account for variation of writing performance shown in SEM. The statistic can not be compared with the above number. EFL writers can not write better with native speakers in some aspects. If they write longer than average, more grammar problems will appear in the context, raters would make an unfavorable judgment.

Thirdly, accuracy, the least predictor in SEM, does not provided the same contributions to other two dimensions. Accuracy focuses finding the mistakes of the composition while complexity paid more attention to choosing words and sentences. It is not difficult to understand in testing, human raters focused more about lexical and syntactic richness in limited time to finish scoring the task especially for L2 writing. As a result, accuracy is the least factor to affect final score probably because of holistic scoring employed in the study.

4.2.3 Discussion on Research Question 3

As for the high proficiency learners and low proficiency learners, how does the understanding of the different dimensions influence the writing performance?

This part use qualitative data to illustrate and confirm what have found in the above part, and probe into the differences in these dimensions between high proficiency learners and low proficiency learners. The sources of data include several subjects' responses to the interview questions and the compositions the subjects have written.

The questions should be probed into thoroughly:

1. Can the findings relating data be tested by the interview?

2. How does interview reveal the importance of form on L2 writing?

3. What factors affect the high proficiency learners and low proficiency learners?

Four students were selected as subjects to do interview. Subjects had been chosen according to the score of the compositions.

Subject 1 and 2 were high proficiency second language student writers whereas Subject 3 and 4 were low proficiency second language student writers. Quantitative data reveal differences in form among these four subjects (see Table 4.12). The high proficiency learners produced second language composition with better content, organization and language than the low proficiency learners. Interview used to probe into the underlying reasons that caused the differences and give insights into understanding of fluency, accuracy and complexity.

Table 4.11 Holistic Score of Writing Performance of Subjects

Subject	**1**	**2**	**3**	**4**
Holistic score	92	90	62	63

Table 4.12 Analytic Score of Form of Subjects

Subject	**1**	**2**	**3**	**4**
Fluency	43	40	25	24
Complexity	40	44	19	15
Accuracy	40	32	28	13

According to Table 4.11 and Table 4.12, it can be seen clearly Subject 1 and Subject 2 had done excellent job about writing in the examination. They got very high score as for holistic score. As for analytic score, subject 1 had achieved high score in three dimensions whereas subject 2 had got score in complexity. Subject 3 and Subject 4 passed the examination. Both of them had got much lower score than Subject 1 and Subject 2.

Before interview the subjects, the researcher had prepared several questions, intending to know more about the writing process and different emphasis. In the interview, the researcher was a good listener while the subjects stated their ideas. Try not to interrupt the subjects and giving more time to them, if they can express themselves as elaborately as possible. It was necessary to ask other questions if more knowledge can be acquired from the subjects. All of the interview procedure was recorded. After the interview, the research listened again and again to know more.

Different subjects gave various kinds of responses to the above questions. Both Subject 1 and Subject 2 knew clearly that this composition is argumentative writing which is composed of argument, evidence and argumentation. They knew clearly that they should make their own statements clearly and convincely. As the second language learners, they knew there are many differences between Chinese writing and English writing. After several years of learning foreign language, they had acquired more knowledge of second language writing.

Interview questions were about the form presented in the essay:

(1) While writing, which aspects do you focus on?

(2) How did you choose words and grammar?

(3) How did you solve the problems when you meet the difficulties of expressing your ideas?

(4) Do you have the habit of paying attention to revising the mistakes in a piece of article? What kind of mistakes would you like to revise?

Subject 1 believed a piece of good writing is made up of elaborate content as well as sophisticated words and sentences. For her, choosing words is not difficult when writing. She is aware of using complicated sentence pattern to express herself. She is confident and she said:

> "I learned writing course in last semester. When I see the topic and the direction of the composition, I know this is argumentative article, I should state my argument clearly at the very beginning. I prefer to mark the statement at the very beginning, although I know topic sentence can be written at the beginning, in the middle or at the end of the passage. I believe that it is helpful to the readers if I state my opinion at the beginning. And I have learned Chinese for so many years, in my opinion, a piece of good writing should be well-organized, referring to content as well as form. Words, sentences are very important in writing. I like remembering the words in daily life and I am aware that the different usage of the same word in different context. I prefer reading magazines and newspapers. From reading, I have accumulated a number of words, so I don not think choosing words is the serious problem for me. I have a good master of grammar, I know clearly how to use emphasized sentence, inverted sentence, attributive clause and adverbial

clause. It is natural for me to express myself when I use English. Of course, there is the exception. If I can not express my ideas freely by the standard English, I will try to use other ways of showing my original idea. As for revising, I just read the whole article as quickly as I can because time is limited. If I am aware of the mistakes of spelling, of course, I will correct it. I have no time to revise the whole sentence."

Subject 2, another high proficiency learner considered vocabulary is a big problem in writing. She stressed that she should spend more time to select the words carefully during L2 writing. She said:

"I had written the outline of the article when I see the title. When I write the article, I focused more attention on word choice. Sometimes, I am puzzled to choose suitable words. I always spend more time to select the appropriate words seriously when writing the composition. I can not control well. Although I keep in mind the similar words, I can not decide as quickly as possible. I believe expressing perfectly can give great impression on the readers and raters. I am expected to be unique. In terms of grammar, I paid more attention to use attributive clause, which is closer to the foreigners' way of thinking. However, form is less important than content. I believe a good article should be of full content and standard English. When I can not convey my ideas because of the language barrier, I will try my best to solve by using clause. As to the last question, I believe I have spent much time in writing process, I have no extra time to revise carefully in the end of testing. I just pay attention to punctuation, paragraphing, and spelling. "

Subject 3 and subject 4 regarded vocabulary and grammar as the biggest problem in second language writing. Subject 3 said:

"I know clearly, words and sentences are very important in essays. Content should be expressed with the help of form. When I write something in English, I often can not find the suitable words to express my ideas even if the simple ones. I am worried a lot when I write in English. I don not like reading and remembering. I have a small sum of words. Most of the time, the ideas I think of can not be translated into English, let alone grammar. I am not good at using

the complicated sentences even the absolute structure and subjunctive mood. I did not learn well in senior middle school. Therefore, I enjoy using the simple sentence in English in order to avoid the mistakes. If I think I can not translate the sentence into English, I will give it up. Generally speaking, I have no time revising the article after finishing writing. I am very nervous in testing and I have no expectation but finish the article in limited time."

Subject 4 said:

"I believe form is very important when do L2 writing, but I can not do well in this aspect. When writing something in the examination, you know, time is limited. I am eager to finish writing, I have no time to take the advanced words into consideration. If there is no time limited, maybe I can do well in choosing words and sentences. Grammar is not my strength, I make sure there is no mistake in the sentence, that is enough. Chinglish is unavoidable for me, but I can do nothing about that. Sometimes I cannot realize that is not standard English unless others told me clearly. If I have time, I can revise the essay from spelling, which is enough. In fact, I don not have time to make revision. When I write the article, I can think more but have difficulty in expressing. At I last I have to write what I can express in English"

These questions are concerning more about form. It is easy for the researcher to understand what they have said and find something in common among different subjects. It is not difficult for the research to conclude that, after so many years of education, especially higher education, most of them show great concern about the big event in our country. As the university students, they focused their attention on job marketing. The topic is appropriate.

Admittedly, all of the subjects concentrate their energy and time on form, language usage. To some extent, they have realized the significance of form to good writing. But they did not emphasize the importance of fluency. To some degree, complexity drew their more attention than fluency and accuracy. Some of them, especially high proficiency learners, do not neglect paragraphing, punctuation, and spelling. While rating, fluency, accuracy and complexity are three indicators of the composition. Teachers primarily concern with fluency, accuracy and complexity, whereas students paid more attention to complexity and accuracy. One possible

reason is that in the current scoring rubric, advanced words and complicated sentence pattern can add more to writing performance. The researcher had prepared the question for giving deeper insights into the subjects' ability of language application. According to their responses, the researcher can arrive at the following conclusions.

First, all of the subjects have realized the importance of form in writing, although most of them can not use the academic term to describe the features of writing. As to this point, there is no difference between high proficiency learners and low proficiency learners. During pre-writing and writing, they always spend more time on language use.

Second, all of the subjects have paid attention to choosing words and sentences. That is to say, they would like to make the article complex and accurate. The difference is that: high proficiency learners are of great awareness of using the appropriate words and special sentence patterns in English writing. Both of the two subjects show confidence of words usage and sentence choosing. Low proficiency learners can not find the suitable words to express their ideas even simple ones. Let alone the sentence. They are used to writing the simple ones. Both high and low proficiency learners consider the words choosing carefully. Different choices in choosing expressing ways are root of the ability of second language knowledge and competence.

Third, in view of the relationship of content and form, they provide different ideas. High proficiency learners consider more about content than form largely due to having a good master of English, while low proficiency learners view it differently. They regard form as the most important one in writing, if she can use the advanced words or complex sentences, she can get higher mark in the examination. This idea is formed by misleading of testing. And this idea presents misunderstanding of scoring criteria, or results from feedback from teachers in training.

Fourth, fluency can be evaluated by length and readability of the composition. There is no doubt that high proficiency learners can write longer compositions than low proficiency learners. The longer the composition is, the more detailed ideas they may show and the higher score they may get for the second language compositions. Different proficiency learners provide different length of composition.

Chapter Five Conclusion

First of all, the major findings of the study are presented in 5.1 and the implications arel listed in 5.2. Then in 5.3 some limitations of the present study can be involved. At last, in 5.4 some suggestions for future study are provided.

5.1 Major Findings

The study has dealt with the relationships between form and writing performance as well as different contributions of three dimensions of form to writing performance. And the study has concerned different effects of form to different groups of second language proficiency.

Using the combination of quantitative and qualitative approaches, from the empirical and descriptive study mentioned in the last two chapters, we can arrive at the following findings clearly.

Firstly, form is positively correlated with the writing performance. From SEM, form can account for 47.61% variance of writing performance, that is to say, writing performance can be predicted by form of the composition to a great degree.

Secondly, as the specific elements of form, three dimensions—fluency, complexity and accuracy—have strong correlation with writing performance. Specifically, according to the statistics in the study, the path coefficient of fluency, complexity and accuracy of writing performance is .66, .72, .29. Complexity is highly correlated with writing performance, whereas accuracy is the least factor concerning writing performance. In other words, complexity can predict more about writing performance. Accuracy can predict the least among three dimensions.

Thirdly, the results of correlation analysis imply that different groups of students show different understanding concerning the contribution of form. High proficiency learners can do better in all sides, while low proficiency learners can not perform well. Fluency, complexity and accuracy are constant factors to predict L2 writing performance, regardless of different levels of student. The higher level the student is, the stronger the complexity accounts for the predictability.

Fourthly, L2 writers have noticed the importance of form, especially the high proficiency learners. High proficiency learners had mastered the knowledge of writing and express the ideas to the fullest use. On the contrary, low proficiency learners can not control well what they have learned during writing process.

During the empirical study, the researcher collected the samples of 243 students. Then, six raters have been involved to offer the holistic and analytic scoring. Based on the collected data, correlation analysis according to SPSS (16.0) and SEM can be applied in the study. The interview of subjects can confirm the results of quantitative analysis. Therefore, the major findings can be perceived.

5.2 Implications

The study has provided theoretical implications as well as practical ones. From theoretical aspect, the implications of the study are as follows:

The study has tested the validity of second language writing through the data analyzed by SPSS (16.0) and using classical SEM. To some extent, it fills in the blank of second language assessment, which has theoretical implications to the relevant study of this subfield.

The study has confirmed the validity of analytic scoring. It can solve the problem of holistic scoring, which bring the low validity and reliability in present. In addition, compared with other analytical scoring, this model can offer three dimensions on form, and these three dimensions can predict writing performance properly, thus it can improve the efficiency and decrease the cost of assessment.

From practical aspect, the present study has provided some practical implications for teaching and learning as well as the testing of L2 writing performance.

In the first place, there is much significance in L2 teaching. According to the traditional writing teaching, content and form are two basic elements constituting a piece of writing.

On the one hand, as the relationship between content and form stated clearly in the study, the teachers should give emphasis to both parties when teaching how to write the high quality article. Namely, some suggestions should be put forward as follows. Firstly, teachers should lay more emphasis on how to understand the topic, how to make the statement clearly, how to use cohesion and coherence, and how to make the argument persuaded. Teachers should

emphasize the importance on different way of thinking and different way of writing between Chinese and English. In writing training, in spite of narrative or argumentative writing, the teacher should evaluate students' writing in the above aspects. Secondly, Teachers should encourage students to improve lexical, grammatical and discourse competence. Something related form such as fluency, accuracy and complexity of the words and sentences should not be neglected. Teachers should also encourage students to use clause and special sentence pattern and passive voice in L2 writing. Try to help students avoid the dilemma, that is to say, students have a lot to write, but it is not easy to express in English. Thirdly, Teachers should provide more argumentative writing to students, and let students write more in training. Read and write as more as possible, which is the only solution to improving writing ability.

On the other hand, according to the specific scoring criteria, the teacher can know the strong points and weaknesses of students in foreign language writing, and then give effective guidance to improve their teaching. The best way to assess students' writing is analytic scoring in L2 writing training. Although it is difficult to manipulate, it is better for the teachers to master more information from students. When students hand in their compositions, it is of great benefit for teachers to adopt this assessment. After scoring the compositions, teachers can analyze each dimension directly, giving different reviews and feedback to each student, which is helpful to improve teaching.

In the second place, the present study has given several suggestions in terms of learning. Firstly, analytical scoring can provide more effective feedback to the test participants from different multi-facet. From the students' point, many weak points should be concerned by students themselves. For example, students must cultivate the awareness of using topic sentence, especially put it at the very beginning of each paragraph, which is helpful for readers to get the main idea as quickly as possible. Another case in point, critical thinking should be cultivated in order to think logically. While organizing the article, the students' logical thinking play an important role. Under this circumstance, they can write the essay united and logically. They can acquire detailed feedback from analytic scoring, which can improve their writing. They know clearly the strong points and weak points of their writing. In the following writing training, they should continue the strength and avoid the weakness.

Secondly, they know clearly that content and form play crucial role in foreign language writing. During writing practice in daily life, they should not only focus on the content of the composition, they should also lay more emphasis on advanced words and complicated sentence patterns. For one thing, in practice, they had better read more original editions to acquire words usage in different context. They should be aware that complicated sentence patterns, passive voice and ellipsis can bring more changes of the composition. When do timed-limited essays, it is better for students to use something unique to express ideas. For another, students read more to increase the material in their mind. They had better think logically consciously or unconsciously. They should cultivate different way of thinking, trying to get closer and closer to the foreigners' way of thinking. As one way of output, writing can display one's way of thinking, which is of great significance for English learners.

Thirdly, as is known to all, in the present examinations in China, holistic scoring is more popular than analytic scoring, and the teachers always place more emphasis on form rather than content during teaching. That is not only one kind of misunderstanding of the criteria of assessment, but also misleading students when they do L2 writing.

In terms of testing L2 writing performance, the present study had confirmed two different methods of scoring, although both of them are of high validity, the researcher consider that analytic scoring has more advantages when used in the daily teaching, whereas holistic scoring can be practiced more in the examination. Analytic scoring should be applied in the extended span.

In sum, both the teachers and students can benefit a lot when they cultivate the awareness of the importance of content and form. Neither of both sides should be neglected.

The findings are helpful for us to improve the way of assessing L2 writing. To some extent, it fills in the blank of second language assessment, which has theoretical implications to the relevant study of this subfield. Meanwhile, it can provide more effective feedback to the test participants from different multi-facet, which is helpful to improve the writing ability.

5.3 Limitations

The study is designed cautiously and carefully at the initial period, however, there are still some weak points.

Firstly, in the experiment, there are only 243 students chosen from the universities in Hebei Province, which can not stand for the whole level of English majors in China. If there are more students coming from different level of university, maybe the quality of writing should be identified. Of course, there is a little change about the data even the result. In fact, it is too difficult for the researcher to find ideal participants who can represent English majors at all levels.

Secondly, it is well known that the compositions can not be well assessed objectively. Although the raters have been trained together at the very beginning, they must have understood the criteria of scoring to the high degree. Inconsistency of rating is still one of the unavoidable factors during the study. In a word, some subjective factors may exist. What's more, reliability can be improved if there are more raters involved in the study, definitely speaking.

Thirdly, in the study, subjects are required to finish one argumentative article in the examination. If necessary, more tasks should be adopted in the future study.

5.4 Suggestions for Future Research

The findings of the study can provide some suggestions for future studies.

Firstly, future study can compare timed-essay writing and untimed-essay writing. The results can be compared to see the independent variables affecting L2 writing performance under two different writing conditions.

Secondly, this study has found that form has contribution to the writing performance. Future studies are needed to investigate both content and form thoroughly about factors in writing process affecting the achievement of L2 writing.

Lastly, further study can investigate different styles of writing according to the same scoring scale, for example, narrative writing or expository one.

Bibliography

Akbar Afghari. 2004. Learner Variables and ESP Test Performance. *IJAL , Vol.7, NO.1*

Atkinson, D. (ed.) 2003. L2 Writing in the Post-process. Era: Introduction. *Journal of Second Language Writing.* 12(1)

Bacha, N. 2001. Writing Evaluation: What can Analytic Versus Holistic Essay Scoring Tell Us? System 29: 371—3.

Bachman L.F., & **Palmer**, A. S. 1996. Language Testing Practice. CUP, 82-101

Bachman, L. F. & **Palmer**, A. S. 1996. *Language Testing in Practice.* Oxford: Oxford University Press, 1-160.

Barrett, P. 2001. Assessing the reliability of rating data.(On line) *Retrieved* April 11, 2005 http:// www.liv.ac.uk/ pbarrett/ ratter.pdf.

Berglund, Y. & Mason, O. 2001. But this formula doesn't mean anything...!?: some reflections on parameters of texts and their significance.(On line) *Retrieved* October 11, 2004 http:// web.bham.ac.uk/o.mason/publications/cl2001/berglund-mason.html

Blok, H., and de **Glopper**, K.1992. *Large Scale Writing Assessment.* In L. Verhoeven and J. H. A. L. De Jong (eds.). The Construct of Language Proficiency. Amsterdam/Philadelphia: John Benjamins pp.101-111

Brown, H.D. 1994. *Teaching by principles: an interactive approach to language pedagogy.* Englewood Cliffs, NJ: Prentice Hall Regents.

Carlson, S. 1988. Relationships of Reasoning and Writing Skills to GRE Analytical Ability Scores. *GRE Board Report* No.84-23P.

Carolyn Boyce and Palena **Neal**. 2006. Conducting In-Depth Interview: A Guide for designing and Conducting In-Depth Interviews for Evaluation Input, Pathfinder International Tool Series.

Casanave, C. 1994. Language Development in Students' Journals. *Journal of Second Language Writing* 3: 179-201

Clement, R. & B. G. **Kruidenier**. 1985. Aptitude Attitude and Motivation in Second Language Proficiency: A test of clement's model. *Journal of Language and Social*

Psychology 4:21-37.

Cohen, A. D. 1994. *Assessing Language Ability in the Classroom. (2nd ed).* Boston, MA: Heinle and Heinle

Cooper, P L. 1984. The Assessment of Writing Ability: a Review of Research. Princeton: NJ: ETS, *GRE Board Research Report*, No 52-15R, 24

Csizer, K. & Z. **Dörnyei**. 2005. The Internal Structure of Language Learning Motivation and its Relationship with Language Choice and Learning Effort. *The Modern Language Journal*. 89(1):19-36

Darlington, R. B. 1968. Multiple Regression in Psychological Research and Practice. *Psychological Bulletin,* 69,161-182

Ely, C. M. 1986. Language Learning Data: A Description and Causal Analysis. *Modern Language Journal* 70: 28-35.

Engber, C. A. 1995. The Relationship of Lexical Proficiency to the Quality of ESL Compositions. *Journal of Second Language Writing*, 4(2), 139-155

Fillmore, C.J. 1979 On Fluency. In Fillmore, C. J., Kempler, D& Wang W. S.-Y. (Eds.), *Individual Differences in Language Ability and Language Behavior*. New York: Academic Press.

Foster, P., & **Skehan**, P. 1996. The Influence of Planning and Task Type on Second Language Performance. *Studies in Second Language Acquisition*, 18, 299-323

Francis, J. C. 1977. Impression and Analytic Marking Methods. *Memeo, MS Aldershot: Associated Examining Board.*

Freedman, S. W. 1979. Why do Teachers Give the Grades They Do? *College Composition*, 32, 365-387

Gardner, R. C., R. N. Lalonde & Pierson. 1983. The Socio-education Model of Second Language Acquisition: An Investigation Using LISEL Causal Modeling. *Journal of Language and Social Psychology* 2: 1-15

Gardner, R. C., R. N. Lalonde, R. Moorcraft & F. T. Evers. 1987. Second Language Attrition: The Role of Motivation and Use. *Journal of Language and Social Psychology* 6: 1-47.

Gardner, R. C. 1988. The Socio-educational Model of Second Language Learning:

Assumptions, Findings and Issues. *Language Learning* 38: 101-26.

Gelderen, A., Schoonen, R., De Glopper, K., Hulstijn, J., Snellings, P., Simis, A., & Stevenson, M. 2003. Roles of linguistic knowledge, metacognitive knowledge and processing speed in L3, L2 and L1 reading comprehension: a structural equation modeling approach. *The International Journal of Bilingualism*, 7(1), 7-25.

Grobe, C. 1981. Syntactic Maturity, Mechanics, and Vocabulary as Predictors of Quality Ratings. *Research in the Teaching of English*, 15, 75-85.

Hamp-Lyons, L. 1990. *Second Language Writing: Assessment Issues*. In B.Koll (eds), Second Language Writing: Research Insights for the Classroom. New York: Cambridge University Press.

Hamp-Lyons, L. 1991. *Pre-text: Task-related Influences on the Writer. In L. Hamp-Lyons (ed.), Assessing Second Language Writing in Academic Contexts*. Norwood, NJ: Ablex. Publishing Corporation

Hamp-Lyons, L. 2001. *Fourth Generation Writing Assessment, in Silva, T., Matsuda, P.K.(eds)* 2001. *On sceond language writing.* London: Lawrence Erlbaum Associates, Publishers

Harley, B., & King, M.L. 1989. Verb Lexis in the Written Compositions of Young L2 Learners. *Studies in Second Language Acquisition*, 11, 415-439

Hayes, J. R., Hatch, J. A. & Silk, C. M. 2000. "Does Holistic Assessment Predict Writing Performance?" *Written Communication.* 17.1 (2000): 3-26.

Heaton, J.B. 1988. *Writing English Language Tests.* London: Longman Group UK Ltd, 135.\

Heaton, J.B. 1990. *Classroom Testing*. London: Longman

Hedgcock, J. & Lefkowitz, N. 1992. Collaborative Oral/aural Revision in Foreign Language Writing Instruction. *Journal of Second Language Writing,* 1 (3)

Homburg, T. J. 1984. Holistic Evaluation of ESL Composition: Can it be Validated Objectively? *TESOL Quarterly* 18, 1 87-107

Hughes, A. 1989. *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Hyland, K. 2003. *Second Language Writing*. Cambridge: Cambridge University Press.

Hyltenstam, K. 1988. Lexical Characteristics of Near-native Second Language Learners of Swedish. *Journal of Multilingual and Multicultural Development*, 9, 67-84.

Isaacson, S. L. 1988, Assessing the Written Product: Qualitative and Quantitative Measures, *Exceptional Children,* (54), 528-534.

Jacobs, H., **Zinkgraf**, S. A., **Wormuth**, D. R., **Hartfiel**, V. F., & **Hughey**, J. B.1981. *Testing ESL composition: A Practical Approach.* Rowley, MA: Newbury House.

Kaplan, R. M., Wolff, S.E., Burstein J., LuC., Rock, D.A., & Kaplan, B. A. 1998. *Scoring Essays Automatically Using Surface Features. Princeton*, NJ: Educational Testing Service.

Kellogg, R. T. 1994. The Psychology of Writing. New York: Oxford University Press.

Kenyon, D. 1998. Approaches to Validation in Language Assessment. In A. Kunnan (Ed.), *Validation in Language Assessment* pp. 1-16. Mahwah, NJ: Erlbaum.

Landauer, T. & Dumais, S. 1997. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, 104.211-140.

Larsen-Freeman, D., & Strom, V. 1977. The Construction of a Second Language Acquisition Index of Development. *Language Learning,* 27(1), 123-134.

Larsen-Freeman, D. 1978. An ESL index of development. *TESOL Quarterly*, 12, 439-448.

Larsen-Freeman, D. 1983. Assessing Global Second Language Proficiency. *In H. W. Seliger & M. Long (Eds.), Classroom-oriented Research in Second Language Acquisition.* Rowley, MA: Newbury House.

Larsen-Freeman, D. 2006 *The Emergence of Complexity Fluency and Accuracy in the Oral and Written Production of Five Chinese Learners of English.* Oxford: Oxford University

Laufer, B. 1994. The Lexical Profile of Second Language Writing. Does it Change Over Time? *RELC journal* 25, 2: 21-33.

Laufer, B. and **Nation**, P. 1995. Lexical Richness in L2 Written Production: Can it be Measured? *Applied Linguistics* 16,3:307-322.

Lee, S. 2005. Facilitating and Inhibiting Factors in English as a Foreign Language Writing Performance: A Model Testing with Structure Equation Modeling Language Learning

Lennon, P. 1990. Investigating Fluency in EFL: A Quantitative Approach. *Language learning,*40,387-417.

Lesson, R .1975. *Fluency and Language Teaching.* London Longman Group Limitec

Matsuda, P.K.(eds). 2001. On Second Language Writing. London: Lawrence Erlbaum Associates

Matsuda, Paul, k. 2003. *Second Language Writing in the Twentieth Century: A Situated Historical Perspectives. Exploring the Dynamics of Second Language Writing. ED., Barbara, Kroll*, Cambridge: Cambridge UP, 2003.

McCurry, N., & McCurry, A. 1992. Writing Assessment for the Twenty-First Century. *Computer Teacher*, 19, 35-37

McNamara, T. 1996. *Measuring Second Language Performance.* Addison Wesley Longman Limited: New York

McNamara, T. F. 2000. *Language testing.* Oxford: Oxford University Press.

Murphy, R. J. L. 1982. A Further Report of Investigations into the Reliability of Marking of G. C. E. Examinations. *British Journal of Educational Psychology*, 52/1, 58-63

Nakamura, Y. 2002. Effectiveness of Paired Rating in the Assessment of English Compositions. *JLTA Journal, 5,* 61-71.

Nakamura, Y. 2004. A Comparison of Holistic and Analytic Scoring Methods in the Assessment of Writing. The Interface between Interlanguage, Pragmatics and assessment: Proceedings of the 3rd annual JALT PAN-SIG Conference. May 22-23

Nation, P. 1989. Improving Speaking Fluency. *System*3, 377-384

Nation, P. 2001. *Learning Vocabulary in Another Language*. Cambridge, Cambridge

Ortega, L. 2003. Syntactic Complexity Measures and Their Relationship to L2 Proficiency: a Research Synthesis of College-level L2 Writing. *Applied Linguistics* 24/4: 492/518

Osborne, J. W. 2000. Prediction in Multiple Regression. Practical Assessment, Research & Evaluation, 7(2).

Page, E. B. 1968. The Use of the Computer in Analyzing Student Essays. *Int'l Rev. Education,* Vol. 14, 1968, pp. 210-225.

Page, E. B. 1995. The Computer Moves into Essay Scoring: Updating the Ancient Text. *Phi Delta Kappan* March, 561-565

Polio, C. 1997. Measures of Linguistic Accuracy in Second Language Writing Research.

Language learning 47:1, March 1997, pp. 101-143

Polio, C. 2001. Research Methodology in Second Language Writing Research. In Silva, T &

Polio, C. 2003. Research on Second Language Writing: an Overview of What We Investigate and How. In B. Kroll (ed) 2003. *Exploring the dynamics of second language writing.* Cambridge: CUP

Prater,D. & **Padia**, W. 1980. Effects of Modes of Discourse in Writing Performance in Grades Four and Six. Paper presented at the annual meeting of the American Educational Research Association, Boston

Purpura, J.E. 1996. An analysis of the Relationships between Test Takers' Cognitive and Metacognitive Strategy Use and Second Language Test Performance. Language Learning, 47, 289-325.

Purpura, J.E. 1998. Investigating the Effects of Strategy Use and Second Language Test Performance with High- and Low-ability Test Takers: A structural equation modeling approach. *Language Testing*, 15, 333-379.

Purves, A. C. 1985. In Search of an Internationally Valid Scheme for Scoring Compositions. *College Composition and Communication.* 35.423-438

Rainmes, A. 1985. What Unskilled ESL Students Do as They Write: A Classroom Study of Composing. *TESOL Quarterly.* 19, 229-258

Reid, S. & **Findlay**, G. 1986. WRITER'S WORKBENCH Analysis of Holistically Scored Essays. *Computers and Compositions.* Volume3, no.2

Robinson, P. 2001. Task Complexity, Task Difficulty, and Task Production: Exploring Interactions in a Componential Framework. *Applied Linguistics*, 22(1), 27-57.

Sasaki, M. 1993. Relationships among Second Language Proficiency Foreign Language Aptitude and Intelligence: A Structural Equation Modeling Approach . *Language Learning* 43: 313-344

Schmidt, R. 1992. Psychological Mechanisms Underlying Second Language Fluency. *Studies in Second Language Acquisition*, 14, 357-385.

Schoonen, R. 2005. Generalizability of Writing Scores: an Application of Structural Equation Modeling. *Language Testing* 22(1): 1-30

Skehan, P. & **Foster**, P. 1997. The Influence of Planning and Post-task Activities on Accuracy and Complexity in Task-based Learning. *Language Teaching Reserch* 1/3

Slotnick, H. 1972. Toward a Theory of ComputerEssay Grading. *Journal of Educational Measurement*, 9, 253-263.

Stemler, S. E. 2004. A Comparison of Consensus, Consistency and Measurement Approaches to Estimating Interrater Reliability. *Practical Assessment, Research & Evaluation,* 9(4). Supercomputer Applications, 6(1), 13-49

Stephen, P. Klein and **Brain** M. Stecher. 1998. Analytic Versus Holistic Scoring of Science Performance Tasks. *Applied Measurement in Education*, 11(2), 121-137.

Stephen Isaacson 1996. "Simple ways to assess the writing skills of students with learning disabilities" On line. http:// www. Readingrockets.org/article

Szmrecsanyi, B 2004. On Operationalizing Syntactic Complexity. JADT 2004: 7es Journees internationals d' Analyze statistique des Fonnees Textuelles

Vacc, N. N.1989. Writing Evaluation: Examining Four Teachers' Holistic and Analytic scores. *The Elementary School Journal 90.*

Vann, R.J. (eds.) 1979. O*ral and Written Ayntactic Relationships in Second Language Learning*. In C. Yorio, K. Perkins, & J. Schachter, On TESOL 79: The learner in focus (pp. 322-329). Washington, D.C: TESOL.

Veal, L. R., & Hudson, S. A. 1983. Direct and Indirect Measures for Large Scale Evaluation of Writing. *Research in the Teaching of English*, 17, 285-296

Weigle, S. C. 1994. Effects of Training on Raters of ESL Compositions. *Language Testing*, 11/2, 197-223

Weigle, S. C. 2002. *Assessing Writing*. Cambridge University Press: Cambridge

Weir, C. 1988. *Communicative Language Testing with Special Reference to English as a Foreign Language.* Exeter: University of Exeter.

Weir, C. J. 1990. *Communicative Language Testing.* NJ: Prentice Hall Regents

White, E.M.1984. Holisticism. *College Composition and Communication* 35(4), 400-409

White, E.M. 1986. Pitfalls in the Testing of Writing. In K.L. Greenberg, H.S. Wiener, & R. A.

White, E. M. 1995. *Teaching and assessing writing.* San Francisco, CA: Jossey-Bass.

Wolfe-Quintero,K., **Inagaki**,S., & **Kim**, H. 1998. *Second Language Development in Writing: Measures of Fluency, Accuracy and Complexity*. Hawaii University Press.

Woodrow, L. 2006. A Model of Adaptive Language Learning. *The Model Language Journal* 90(3): 297-319

Wray, A. 2002. *Formulaic Language and the Lexicon*. Cambridge: CUP

Yashima, T. 2002.Willingness to Communicate in a Second Language: The Japanese EFL Context. T*he Modern Language Journal*. 86(1):54-56

Yashima, T., L. Zenuk-Nishide& K. Shimizu. 2004. The Influence of Attitudes and Affect on Willingness to Communicate and Second Language Communication. *Language Learning* 54(1):119-152

Yong-Won Lee, Claudia **Gentile** & Robert **Kantor**. 2008. *Analytic Scoring of TOEFL CBT Essays: Scores from Humans and E-raters*. TOEFL-RR-81

韩宝成，2006，结构方程模型及其在语言测试中的应用。《现代外语》2006年第1期，78-86页。

侯杰泰、温忠麟、成子娟，2004，《结构方程模型及其应用》。北京：教育科学出版社。

梁茂成， 2005， 中国学生英语作文自动评分模型的构建。 南京：南京大学博士论文。

罗娟，2007， 二语写作测试中的评分量表研究。《中国校外教育》2007 年第 9 期

马广惠，1998，中国大学生的母语写作能力、英语分项能力对英语写作能力的系统影响。南京：南京大学博士论文。

马广惠，2004，《影响二语写作的语言因素研究》。南京：河海大学出版社。

秦晓晴、文秋芳，2007，《中国大学生英语写作能力发展规律与特点研究》。北京：中国社会科学出版社。

王立非，2002，作文语篇知识与课外非命题写作：影响英语写作水平因素的研究。《外语界》2002年第4期，42-49页。

王立非、鲍贵，2003，应用语言学研究的路径分析方法：原理与应用评介。《现代外语》2003年第4期，403-409页。

王立非，2004，《汉语语言能力向英语写作迁移的路径与理据》。西安：陕西师范大学出版社。

王立非，2005，我国英语写作实证研究：现状与思考。《中国外语》 2005 年第 1 期。

王立非，2007，《计算机辅助第二语言研究方法与应用》。 北京：外语教学与研究出版社。

王文宇、文秋芳，2002，母语思维与二语写作——大学生英语写作过程研究。《解放军外国语学院学报》2002年第4期，64-67页。

王文宇，2002，母语思维与外语作文分项成绩之间的关系。《外语与外语教学》 2002年第10期，17-20页。

王文宇、王立非，2004，二语写作研究：十年回顾与展望。《外语界》2004年第3期，51-58页。

温忠麟、侯杰泰、马什赫伯特 2004，结构方程模型检验：拟合指数与卡方准则，《心理学报》2004年第36期，186-194页。

文秋芳，2007，作文内容的构念效度研究—运用结构方程模型软件AMOS5的尝试。《外语研究》2007年第3期，66-71页。

吴红云、刘润清 2004，写作元认知结构方程模型研究。《现代外语》 2007年第4期，370-378页。

吴明隆，2009，《结构方程模型——AMOS 的操作与应用》。重庆：重庆大学出版社。

辛平，2007，基于语言能力构想的作文评分标准及其可操作性研究。《暨南大学华文学院学报》 2007年第3期，19-24页。

修旭东、肖德法 2006，英语写作策略、八级写作认知过程及成绩关系的结构方程模型研究。《外语教学与研究》 2006年第6期，460-465页。

姚琴宜，2008，作文整体评分与分项评分结果的质量对比。《外语研究》2008 年第 5 期。

邹申，2005，《语言测试》。上海：上海外语教育出版社。

Appendix Ⅰ 二语写作评估标准（内容层面）

		一等	二等	三等	四等	五等
作文切题性（50分）	中心论点、分论点与作文要求的相关度	中心论点完全符合作文要求，对大学生做村官现象发表个人观点与看法，与题目相关度高； 分论点与中心论点保持一致，相关度高； 整体论述，包括结论部分，与中心论点及分论点具有较强的相关性。	中心论点与主题相关性较高； 分论点与中心论点保持一致； 整体论述符合作文要求，具有比较强的相关性； 结论部分与中心论点、分论点相关性较强。	中心论点与主题基本保持相关； 分论点与作文要求保持一致，大部分分论点与中心论点一致，个别不一致； 或分论点与中心论点一致，但展开论述时与中心论点相关度降低； 或中心论点不太明确，但分论点与作文要求相一致。 结论能够与中心论点、分论点基本保持一致。	中心论点与主题相关性一般； 分论点与作文要求不完全一致； 或某些分论点与中心论点一致，某些不一致。	中心论点与主题相关性较差，或不相关； 分论点与中心论点不一致。
	分 值	50-41	40-31	30-21	20-11	10-1
	备注（关注点）	1.中心论点与作文要求的相关度；2.分论点与中心论点的相关度；3.论据与分论点的相关度；4.整体论述与作文要求相关度。				
论点明确性（50分）	中心论点与分论点的清晰度	中心论点清晰； 分论点清晰。	中心论点比较清晰； 大部分分论点清晰。	中心论点基本清晰； 大部分分论点比较清晰，个别不清晰。	中心论点基本清晰； 大部分分论点不太清晰。	中心论点不清晰； 分论点不清晰。
	分 值	50-41	40-31	30-21	20-11	10-1
	备注（关注点）	1.是否有篇章句(thesis statement)，由其表述的中心论点是否清晰(state clearly)； 2.是否有主题句(topic sentence)，由其表述的分论点是否清晰。				

篇章连贯性（50分）	分论点间及分论点内部句子间关系的层次性及条理性	分论点与中心论点间：保持一致，具有强的连贯性。 分论点间：各分论点间的关系符合逻辑，具有层次性与条理性。 分论点内部：句子所蕴涵的信息具有语义上强的连贯性。 结论与中心论点、分论点间：它们之间相呼应，尤其对分论点做出合理的概括性总结。	分论点与中心论点间：保持一致，具有比较强的连贯性。 分论点间：各分论点间的关系比较符合逻辑，具有较好的层次性与条理性。 分论点内部：句子所蕴涵的信息具有语义上的较强的连贯性。 结论与中心论点、分论点间：它们之间相呼应，尤其对分论点做出比较合理的概括性总结。	分论点与中心论点间：基本保持一致，连贯性一般。 分论点间：各分论点间的关系基本符合逻辑，具有一定的层次性与条理性。 分论点内部：句子所蕴涵的信息基本具有语义上的连贯。 结论与中心论点、分论点间：能够做到互相呼应，能够对分论点做出一般性的概括总结。	分论点间：分论点间关系基本清楚，但需要读者花气力才能读懂。 分论点内部：句子间连贯性欠缺，所蕴涵信息欠关联。	分论点间：分论点间关系不清楚或无必然联系。 分论点内部：句子间连贯较差，所蕴涵信息没有太大关系。
	分 值	50-41	40-31	30-21	20-11	10-1
	备注（关注点）	1.beginning、body、conclusion 间的连贯性；2. 分论点间是否具有层次性、条理性及逻辑性；3.句子间是否连贯，是否具有逻辑性。				
说理透彻性（50分）	分论点对中心论点的说理充分性；证据对分论点的论证充分性	分论点对中心论点的说理充分性：分论点对中心论点支持作用强，论证充分。 论据对分论点的论证充分性：论述逻辑性强，论据较多，对论点有很好的支持作用。	分论点对中心论点的说理充分性：分论点对中心论点支持作用较强，论证基本充分。 论据对分论点的论证充分性：分论点论述逻辑性强，有举例；或分论点论述逻辑性较强，大部分分论点举例具体。	分论点对中心论点的说理充分性：分论点对中心论点支持作用一般。 论据对分论点的论证充分性：分论点论述清楚，但例证一般。	分论点对中心论点的说理充分性：分论点对中心论点的支持作用欠充分。 论据对分论点的论证充分性：分论点论述比较清楚，举例较少。	分论点对中心论点的说理充分性：分论点对中心论点起不到支持作用。 论据对分论点论证充分性：说理不清楚，无实例或实例不恰当。
	分值	50-41	40-31	30-21	20-11	10-1
	备注（关注点）	1.分论点对中心论点的支持作用是否明显，论证是否充分；2.论据对分论点的支持作用是否明显，论证是否充分。				

注：若不切题，除切题性分值低外，其它各维度得分不能高于及格分。

Appendix II 二语写作评估标准（形式层面）

流利性（50 分）	写作的流利度	表达行文流畅、过渡自然；字数达到要求。	行文比较流畅、过渡比较自然；字数达到要求	行文比较流畅、自然过渡一般；字数基本达到要求	表达不够流畅，自然过渡欠佳；字数达不到要求	语言组织有较大的随意性；字数不够
	分 值	50-41	40-31	30-21	20-11	10-1
	备注（关注点）	1. 阅读流利性；2. 文章长度。				
复杂性（50 分）	词汇复杂性（25 分）	词汇的多样性与丰富性。				
	分 值	起点 8 分，每 1 处加 2.5 分，满分 25 分。				
	句型复杂性（25 分）	句型复杂性、恰当性（内嵌句定语从句、状语从句、倒装句、并列句、反问句等）。				
	分 值	起点 8 分，每 1 处加 2.5 分，满分 25 分。				
	备注	注：口语化句子降一档次。				
准确性（50 分）	大错（25 分）	主要从句子结构入手，如缺少主语/谓语、非谓语动词错误、主谓不一致、时态错误、搭配不当、中式英语、不知所云的句子等。				
	分 值	1 个大错减 2 分。				
	小 错（25 分）	拼写、标点符号、第三人称单数、名词单复数、冠词、问句缺助动词、代词指代不明等错误。				
	分 值	1 个小错减 1 分。				

Appendix Ⅲ TEM4 作文评分标准

14-15 分 内容切题，完整，条理清楚，文章结构严谨，语法正确，语言通顺恰当，句式 用词富有变化，有"闪光点"，基本无语言错误.

11-13 分 内容切题，完整，条理清楚，文章结构严谨，语法正确，语言通顺恰当，少量语法错误

8-10 分 内容基本切题，完整，条理基本清楚，文章结构基本严谨，语法基本正确，语言基本通顺恰当.少量严重错误，一些词使用不当.

5-7 分 内容基本切题，完整，条理不够清楚，较明显的母语痕迹.较多语言错误，许多词使用不当.

2-4 分 内容偏题，不完整，思路混乱，语句不完整，只有少数句子可以理解，词汇拼写错误严重.

Appendix Ⅳ Some Compositions for Analysis

Sample 1

Under the pressure of the economic crisis, college graduates are confronted with a tough situation to find a job. In order to make a living, many of them tend to choose to work as village officials in the countryside. This action has caused hot debate in the society. Many people maintain that it is a good opportunity for the graduates to have some practical experience, and also it is a golden chance for the village to gain some fresh blood. Actually, superficial benefits cannot cover up the deep disadvantages. Working as village officials do harm than good to the graduates, the village and the whole society.

Firstly, tending to be the village officials cannot benefit the college graduates. Actually the purpose of the graduates to be the village officials is not as pure as it appears. Many of the students just take it as an opportunity to find a job which can make them superior in taking exams as civil official. Driven by this motivation, instead of sacrificing themselves in servicing other people, the graduates spare no time in developing the countryside. In this sense, even if they become village official successfully, they themselves cannot benefit from this choice.

Secondly, it adds extra burden to the village and disturbs the normal order of the village. Since the graduates have some special skills, they often require being equipped with advanced facilities, for instance, computers, type writers, etc. All of these will add expense of the village. Besides the expense, the efficiency of working is another serious problem. Owning little practical knowledge, the graduates know little about the method of dealing with the complicated relationship existing in the process of their working. Sometimes, they cannot solve the problems appropriately and efficiently. In this sense, college graduates recruiting as village officials is not a good choice for the villages.

Last but not the least, this choice does harm to the whole society's fairness. In the recruitment, interview occupies a large proportion. However, during the interview many people who are excellent enough find that they are kicked out clxxx with no reason. This is because other people take the interview as an opportunity to give money to the judges. If the

situation continue to go like this, the fairness of our society will not exist anymore. The purpose of our country to propose this policy can never be achieved. Taking all these analysis above into consideration, it's not difficult to find that despite some benefits, the policy of recruiting college graduates as village officials is far from being called a good choice for the students themselves, the village, and our society.

Sample 2

More and more college graduates choose to work as village officials. Some people hold the view that it is a waste of resources because college graduates can do more useful things, while others are in favor of this action. As far as I'm concerned, it is of great significance to both the society and graduates.

First of all, college students spend almost more than 15 years in different levels of schools and have got no chance to expose themselves to the real world. Working as village officials is a golden opportunity for them to accumulate experience while dealing with different kinds of people and complicated things.

Second, college graduates work as village officials plays an important role in shaping the village and injecting fresh element in the relatively backward villages. College graduates are young people who not only have rich knowledge but also a host of brand new ideas towards life and things. By working as working village officials, they can bring knowledge and new management idea to villagers and thus changing the situation in many villages.

Thirdly, college graduates work as village officials is a way of reducing the pressure on employment. In the previous years, college graduates are held high esteem so they only pin their hopes on finding a decent job in big cities. While now amid the global financial crisis, more people choose to start from the grassroot. They just want to find a job then they can choose one if they have gained certain experience.

In spite of all the merits metioned above, there still exists some disadvantages. Some college graduates just treat this kind of experience as a springboard from which they can jump to a higher lever, so they do not work heat and soul to serve people but put their own interests before the public.

In all, it is a kind of action that worth advocating, for it can enhance villagers' overall knowledge, improve the graduates themselves and at the same time, reduce the pressure on employment to some extent.

Sample 3

In the current social situation, people's educational level is higher and higher. Now the villages are on their way of changing their image of being poor, vulgar and undeveloping. Thus, more and more intelligent people are required to infuse fresh blood into the village. Then, college graduates working as village officials have become popular. It follows the social laws and it does benefit a lot to the social situation of job hunting, the villages and the graduates themselves.

The first benefit it produces is that it will light the social tension of job hunting. With the dramatic increase of college graduates, job hunting is a headache to many graduates. There are not so many companies meeting so many students. This makes many graduates out of work. However, at the same time, villages are places that no one wants to go because of its poverty and inconvenience. So if many college graduate like to go there to find a job, the job position will not e as crowd as now. This will reduce the pressure of job hunting in the city.

Secondly, college graduates working as village officials will bring benefits to the village. Villages are undeveloping. So no one wants to go there. This has become a bad circulation. If those college graduates are willing to give a hand to the construction of the countryside, it would be a great opportunity for the villages to change their original poor condition. The process of urbanization will not far away be realized.

Finally, college students working as village officials are good to themselves. They may get rid of the air they have as a citier. It is also a exercise for them to mould a great character. They can get exercise in the countryside. Going to the village is good for the development of the graduates' phycology.

Ultimately, the phenomenon that college graduates working as village officials is an opportunity to change the situation of job-hunting, promote village development, and graduate growing-up. If they perform well, our country will become more prosperous and bright.

Sample 4

In recent years, more and more college graduates choose to become village officials. Everything has two sides, including this policy. To be village officials is a two-sided sword, too. Many people think that there are advantages and disadvantages for the graduates to be village officials.

Firstly, someone support the idea that it is beneficial to the group of village officials. (1) College graduates working as village officials can improve the whole quality of village officials. College graduates all accepts good education and training. They own higher cultural quality and insight. Absorbing some graduates to work, the village construction may receive good impact. (2) A large number of college students work as village officials, which can light the working pressure of our country. In the countryside, they can practice their own ability to deal with things. Gradually, graduates can put what they have learned in college into practice.

Although there are so many advantages, some opposite ideas appear at the same time. Some critics tend to say it does not benefit to the whole society and individuals. Their reason are as follows: (1) Graduates have no practical experience to deal with what village officials can. The jobs refer to the policy's carrying out and the steadity of the whole village construction. Graduates are too young to treat the relationship between government and villagers. (2) It has become a kind of blind fashion to attend the exams of village officials. Some graduates cannot understand the meaning of this piece of policy. They just aim to get high salary and stable jobs. What is worse, some of them may take some mean measures to arrive the aim.

This affair has two sides, but as far as I am concerned, the advantages overweigh the disadvantages. It is beneficial for the society and every individual. It's a fashionable policy to follow the change of this age.

Acknowledgments

First and foremost, I would like to convey my sincere appreciation to my supervisor, Professor Gao Guixiang, who kindly gave me constructive suggestions and powerful guidance. She is very responsible, patient and encouraging. From every discussion with Prof. Gao, I have been inspired greatly.

Secondly, special thanks to Professor Pan Bingxin, Professor Zheng Lihua, ProfessorTang Liping, and Professor Gao Zhihuai, for their valuable criticism and helpful suggestions.

Thirdly, I deliver sincere appreciation to my colleague, Associate Professor Gao Xiao, who helped me learn a lot about SEM and gave me valuable guidance concerning research methodology and the statistical tests used in the present study. Meanwhile, thanks to my colleagues who devoted a lot during rating procedure. The study would not have been possible without their active participation.

Finally, I express my grateful to the cooperative students who have participated in the study.

Academic Achievements

文章名称	发表刊物	刊发时间	刊发级别	署名次序
对高职经贸英语专业口语教学的几点思考	科技信息	2006-2	省级	1
英文电影在英语网络自主学习中的应用	电影文学	2008-2	核心	1
基于模因论的网络语言	湖北经济学院学报人文社科版	2008-5	省级	1
模因论与互文性表现力探析	作家	2009-7	核心	1