

外国语言文学学术论丛
Academic Monographs on
Foreign Languages and Literature

英文教师写作能力与 写作评价决策

一项中国内地的探索性研究

刘 力 著

**Writing Proficiency of EFL Teachers and
Scoring Judgment**

An Exploratory Study in Mainland China

 中国人民大学出版社

外国语言文学学术论丛

英文教师写作能力与 写作评价决策： 一项中国内地的探索性研究

Writing Proficiency of EFL Teachers and Scoring
Judgment: An Exploratory Study in Mainland China

刘 力 著

本书受到中国人民大学“统筹支持一流大学和一流学科建设”项目经费资助

中国人民大学出版社
·北京·

图书在版编目(CIP)数据

英文教师写作能力与写作评价决策：一项中国内地的探索性研究：英文 / 刘力著.
—北京：中国人民大学出版社，2016.10
(外国语言文学学术论丛)
ISBN 978-7-300-23564-6

I. ① 英… II. ① 刘… III. ① 英语—写作—教学研究—英文 IV. ① H315

中国版本图书馆 CIP 数据核字（2016）第 261931 号

外国语言文学学术论丛
英文教师写作能力与写作评价决策：一项中国内地的探索性研究
刘 力 著
Yingwen Jiaoshi Xiezuo Nengli yu Xiezuo Pingjia Juece: Yixiang Zhongguo Neidi de
Tansuoxing Yanjiu

出版发行	中国人民大学出版社		
社 址	北京中关村大街 31 号	邮政编码	100080
电 话	010-62511242（总编室） 010-82501766（邮购部） 010-62515195（发行公司）		010-62511770（质管部） 010-62514148（门市部） 010-62515275（盗版举报）
网 址	http://www.crup.com.cn http://www.ttrnet.com（人大教研网）		
经 销	新华书店		
印 刷	北京易丰印捷科技股份有限公司		
规 格	148 mm × 210 mm 32 开本	版 次	2016 年 10 月第 1 版
印 张	8.625	印 次	2016 年 10 月第 1 次印刷
字 数	326 000	定 价	28.00 元

版权所有 侵权必究 印装差错 负责调换

Preface

This book would not have been possible without the generous help and consistent support from a number of people. First and foremost, I'm greatly indebted to my parents. Their unwavering love, support and patience have allowed me to come this far. I would also like to thank Deng Pan, my husband, for his incredible patience, care and good cheer all the time. He has been a constant believer in me, who always encourages me when things looked their bleakest and reminds me that there is, indeed, life beyond research.

Specially, I would like to express my gratitude to Prof. Barley Mak from The Chinese University of Hong Kong and Prof. David Coniam from The Education University of Hong Kong. Discussions with them have been my intellectual challenge and enjoyment. Barley's standards for scholarship have made me the researcher I'm today and given me the confidence to continue in academia. I'd like to express my special appreciation to Dave for his patient support and professional guidance at all stages of my research as well as the generous time spent reading my drafts. Also, I'd like to convey my heartfelt thanks to Prof. Yan Jin from Shanghai Jiao Tong University for providing

me with constructive comments on the draft and the suggestions regarding my future academic development.

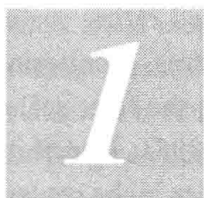
In addition, I am indebted to a number of people for their advice and assistance in various ways. They are Prof. Yan Zhou from Beijing Foreign Studies University, Prof. Zheng Zhang from Beijing Normal University, Prof. Li Lin and Dr. Linlin Cui from Capital Normal University. I'm also grateful to the pre-service EFL teachers and the raters who gave so generously of their time in taking part in two phases of the research.

Friends play a really important role in accompanying me during my journey of writing. Specially, I'm grateful to Dr. Tan Jin from Sun Yat-sen University for his professional assistance in reviewing and commenting on my work and the wonderful morning teas. I am hoping the collaboration can still be continued in the future.

To all these people, I dedicate this book.

Contents

Chapter 1	Introduction	1
Chapter 2	Writing Proficiency of Pre-service EFL Teachers	15
Chapter 3	Rater Judgments in Writing Assessment	47
Chapter 4	Study One: Writing Proficiency	83
Chapter 5	Study Two: Scoring Judgments	130
Chapter 6	General Discussion and Conclusion ...	223
Appendices.....		241
References		252



Chapter 1

Introduction

1.1 Context of the research

1.1.1 Rater variability in performance assessment

As a direct measure of learners' communicative language ability, performance assessment is commonly espoused for its close link between the assessment situation and authentic language use and is often taken for granted to enhance the validity of inference drawn from the scores (Bachman et al., 1995; Lynch & McNamara, 1998). It has therefore been increasingly involved as a compulsory or optional part in many large-scale language assessments worldwide.

However, the elicitation of complex responses from candidates inevitably calls for raters to make evaluative judgments on the effectiveness of candidate performance or the degree of mastery of the underlying construct that the assessment intends to measure. Variability in scores has been regarded as a "measurement error" that lowers the reliability and, hence the validity of the test (Hout, 2002; Ruth & Murphy, 1988). Research has focused on increasing the score consistency across raters (intra-/inter-rater reliability), writing tasks and occasions (test-retest reliability) by controlling or reducing the variability due to rater factors. Various procedures therefore have been

proposed including training raters, using standardized rating scales to direct the raters to look for the same writing features, monitoring raters regularly to check their consistency in applying the rater scales and adopting double marking, etc. (e.g., Jacobs et al., 1981; Underhill, 1982). However, research findings from numerous studies in the context of performance assessment indicate that even after principled rater training or standardization, raters still exhibit considerable variability or idiosyncrasies in the ratings they award (Lumely & McNamara, 1995; Weigle, 1998). Rater factors have therefore been deemed as one of the significant sources of variability that researchers have explored and taken into account when interpreting and using test scores to make valid and fair decisions (Cumming, 1997; Hout, 2002; Schoonen, 2005). It is reasonable to assume that raters mediate between the candidate performance and the final score with their internalized criteria and specific approach in implementing these criteria, to determine the meaningfulness of the score and the appropriateness of inferences made from the results.

Two lines of research have been identified to investigate rater variability by exploring various underlying factors which lead to the observed variations. One line of studies has been concerned with variability introduced by raters' judgment through statistical analysis of the scores awarded by the raters. The other has focused on raters' rationales for their scoring decisions. Rather than focusing on the final scores, this line of research perceives raters as decision-makers who go through different thought processes to arrive at the final scores.

The first line of research has focused on the statistical modeling of rater effects. The most commonly used statistical methods in detecting and measuring rater variability include inter-/intra-rater reliability indices by Classical Test Theory (CTT), estimation of variance component related with the whole rater facet by Generalizability Theory (G-theory) and calibration of individual rater's rating patterns by Many-Facet Rasch Model (MFRM). By operationalizing rater variability in different

ways, these methods provide different statistical indices depicting raters' ratings from different perspectives. Though these more sophisticated statistical methods such as G-theory and MFRM enable researchers to investigate rater factors in a more systematic manner than the correlation coefficients in the CTT, they leave the complexity and interaction in the rating process unexplored. Many researchers therefore have called for more in-depth investigations into the rating process articulated in the future agenda of their research (e.g., Weigle, 1998; Eckes, 2005).

The other line of research is therefore devoted to exploring how raters arrive at their scoring decisions, with the aim of finding out the underlying reasons leading to the individual scoring judgments among raters. Some of the studies have drawn upon indirect evidence to infer what writing features raters attend to (Cumming et al., 2006; Eckes, 2005; Homburg, 1984; Jerkins & Parra, 2003; Laufer & Nation, 1995). These endeavors help to extract salient features in candidate performance which influence raters' decision-making and therefore provide useful information for the validation or development of rating scale. There are also other studies which have employed intro-/retrospective verbal protocols as a direct evidence of raters' thought processes in their decision-making. Some of these studies have focused on describing the similarities and differences in raters' rating focus and their reading styles as well as the strategies employed to acquire and process the information (e.g., Vaughan, 1991; Milanovic et al., 1996; Sakyi, 2003; Cumming et al., 2002, 2003; Lumely, 2002, 2005). Given that the nature of these studies is exploratory; their findings are mixed due to the specific assessment context and different rater groups investigated, more importantly, to the complex nature of the rating process.

Though small in number, these studies have revealed important sources of variance among raters with different personal, cultural and professional backgrounds. At the same time, these studies still fall short in terms of providing a comprehensive account of the mechanism of how these factors lead to rater variability in both their scoring outcomes

and processes. A mixed-method approach examining both raters' rating outcomes and processes can therefore enhance understanding of raters' scoring judgment and explore how it can be related to the factors influencing their rating.

1.1.2 Pre-service EFL teachers and writing assessment

1.1.2.1 *Writing proficiency of pre-service EFL teachers*

Language proficiency has long been recognized as one of the most essential characteristics of a good language teacher (Lange, 1990). This recognition has given rise to concerns about language teachers' proficiency, particular for English-as-a-foreign-language (EFL) teachers (Arva & Medgyes, 2000; Coniam & Falvey, 1996, 2000, 2001, 2007; Elder, 2001). Nowadays, EFL teachers are faced with the challenge of achieving appropriate levels of target language proficiency for delivering effective instruction and carrying out their professional activities. In the context of Asia-Pacific countries, Nunan (2003) suggests that English language proficiency for many teachers is not sufficient enough to provide learners with the rich input needed for successful foreign language acquisition. Apart from the important role that language proficiency plays for students, it has been argued that the language proficiency of teachers is too often overlooked (Johnson, 1990; Richards, 1998).

Of the four macro-skills related to language proficiency, it has been argued that learning to write in a second language (L2) is far more challenging than learning to listen to, to speak or to read a foreign language (Bell & Burnaby, 1984; Bialystok, 1978; Nunan, 1989). Writing requires coordinating a complex and multifaceted set of skills and learning these skills requires careful instruction and guidance from teachers who are competent and confident in their writing ability (Ochsner & Fowler, 2004). For a writing class, teachers have to

meaningfully respond to and critically evaluate students' written works such as the ones produced under traditional writing tests, which are then scored on some sort of numerical scale (Hamp-Lyons, 1991) or other informal assessment activities such as portfolios or take-home writing assignments. It can be seen that teachers' capabilities for evaluating writing and their competence to provide feedback to students are closely tied to their ability to judge varying levels of writing quality and use these judgments in providing their students with diagnostic feedback (Dappen, Isernhagen & Anderson, 2008).

However, teachers' writing performances are far from satisfaction. Compared the results of teachers taking *Language Proficiency Assessment for Teachers of English* (LPAT) on different papers from 2001 to 2011, the scores on the Writing papers are the weakest (Education Bureau, 2011; Lin, 2007). Current research on writing proficiency of teacher candidate is scant at best and no data on the writing proficiency level of pre-service EFL teachers can be attained. In addition, compared with the bulk of studies on students in relationship to assessment, relatively few studies have examined empirically the evaluation criteria and assessment practices of EFL teachers in the classroom context (Xu & Liu, 2009).

1.1.2.2 Pre-service EFL teachers as raters

Language teachers, novice or experienced, are usually involved as raters in various language assessments. In the context of large-scale writing assessments, a series of studies have been conducted to explore the influence of rater backgrounds on rating; in most cases, however, only differences between judgments and behaviors of expert/novice or experience/less-experienced raters have been explored (Cumming, 1990; Hout, 1993; Wolfe & Kao, 1996). The influence of raters' language proficiency has emerged most evidently in the contrasting studies of native speaker (NS) and nonnative-speaker (NNS) raters of EFL writing. However, this group of research has yielded ambiguous and inconclusive

findings (e.g., Brown, 1995; Connor-Linton, 1995; Fayer & Krasinski, 1987; O'Loughlin, 1994; Santos, 1988; Shi, 2001).

In the context of classroom assessment, assessing student performance is one of the most critical aspects of the job of a teacher. Research shows that teachers can spend as much as a third to one half of their professional time involved in assessment or assessment related activities (Cheng, 2001). Studies on EFL teachers' assessments of student writing have been conducted to compare with those of native English teachers (Connor-Linton, 1995b; Hamp-Lyons & Zhang, 2001; Kobayashi, 1992; Kobayashi & Rinnert, 1996; Santos, 1988). Some studies have focused on English teachers at tertiary levels assessing heterogeneous EFL students (Cumming, 1990; Brown, 1991; Hamp-Lyons, 1989; Santos, 1988; Vaughan, 1991) or on a homogeneous group of students (Hamp-Lyons & Zhang, 2001; Kobayashi, 1992; Connor-Linton, 1995; Kobayashi & Rinnert, 1996). However, the writing proficiency of EFL teachers has not been investigated as a factor influencing their judgments of student writing. Furthermore, studies investigating the role of teachers' writing proficiency in writing assessment are yet to be conducted in the Chinese context, which is arguably an influential one due to the variety of English and the densely populated country (Berns, 2005).

Typically, in Mainland China, many EFL teachers, in-service or pre-service, perceive the processes of writing assessment as vague and beyond their control, affecting their writing assessment practices (Sheng, 2009). There has been little systematic experience in pre-service teacher education or teacher professional development programs to prepare EFL teachers in writing assessment (Xu & Liu, 2009). A common result is that pre-service EFL teachers enter the profession without any formal training in assessing student writing, rendering it unclear as to whether they will be able to provide quality assessment in the future. It is possible that various factors might influence their judgments of student writing; for example, their own experiences with writing might

be uneven and even negative, resulting in ineffective writing assessment (Bruning & Horn, 2000). The need therefore exists to examine how pre-service EFL teachers make judgments about student writing. Though most pre-service EFL teachers—through their own writing and reading those of others—might be capable of making general judgments about writing, assessing and analyzing student writing at a micro level is a complex and challenging task for them. The explicit connection between pre-service EFL teachers' writing proficiency and classroom assessment practices needs to be established.

To this end, the current book seeks to explore the relationship between writing proficiency of pre-service EFL teachers and their judgments of student writing. Furthermore, no previous study has attempted to employ a mixed-methods approach to investigate the relationship between raters' writing proficiency and their assignment of scores.

1.2 Research questions

The current research combines both quantitative and qualitative methods to investigate the relationship between pre-service EFL teachers' writing proficiency and their scoring judgments of student scripts. Based on previous literature (e.g., Cumming et al., 2001, 2002; Lumley, 2002, 2005; Milanovic et al., 1996, etc.), the writing proficiency is operationalized as the writing features inherent in writing performances of the pre-service EFL teachers. Scoring judgment is defined as pre-service EFL teachers' rating focuses (aspects of writing raters attend to while rating student scripts) and their scoring behaviors. Specifically, two major research questions have been formulated to guide the data collection and analysis procedures.

Research Question 1: How does the writing proficiency of pre-service EFL teachers vary in terms of the discourse features inherent in their writing performance?

Research Question 1-1: To what extent does the writing performance of pre-service EFL teachers differ at various levels of writing proficiency?

Research Question 1-2: In what ways does the writing performance of pre-service EFL teachers differ as characterized by the discourse features?

Research Question 2: How do pre-service EFL teachers at different levels of writing proficiency differ in their scoring judgments of student writing?

Research Question 2-1: How do pre-service EFL teachers at different levels of writing proficiency differ in their rating focuses in the rating process?

Research Question 2-2: How do pre-service EFL teachers at different levels of writing proficiency differ in their scoring behaviors in the rating process?

Correspondingly, two related studies have been designed to address the two research questions. These two studies provide complementary perspectives on pre-service EFL teachers' judgments of student writing: Study One empirically examines actual writing performances of pre-service EFL teachers at different levels of writing proficiency, while Study Two cognitively explores the teachers' scoring judgments, as reflected through the scoring processes. Together, the two studies can provide empirical evidence on the influence of writing proficiency of pre-service EFL teachers in their judgments of student writing scripts.

1.3 Overview of research design

The objective of the study is to explore the relationship between writing proficiency of pre-service EFL teachers and their judgments of student writing. In addressing the major research questions, a mixed-methods design has been adopted with two related empirical studies.

1.3.1 Study One: Writing proficiency of pre-service EFL teachers

Study One was designed to examine differences of writing proficiency of pre-service EFL teachers characterized by discourse features inherent in their writing performances. The discourse analytic approach to writing performances was adopted to provide empirical evidences underlying expert judgments of teachers' writing proficiency levels. Nineteen discourse features were identified based on a systematic document analysis on standards/requirements on writing proficiency of EFL teachers in major English-speaking countries and non-English-speaking regions and countries. Measures that have been used in a range of previous studies and produced reliable, meaningful results, and that have clear theoretical justifications were chosen to operationalize the selected features. Eighty-one pre-service EFL teachers enrolled in one language education program in one key normal university in Mainland China participate in the study. A writing test was designed to elicit their actual writing performances, which were later assessed by language testing experts into three proficiency levels.

Analysis of Variance (ANOVA) was employed and eleven out of nineteen features covering all discourse categories investigated were

able to successfully differentiate between writing performances of pre-service EFL teachers at three assessed levels. In general, the findings of Study One were in line with previous empirical L2 writing studies, providing a detailed picture of writing performances of pre-service EFL teachers.

1.3.2 Study Two: Scoring judgments of pre-service EFL teachers

To address Research Question 2, Study Two investigated the scoring processes of pre-service EFL teachers to determine how their judgments teachers related to their levels of writing proficiency. Think-aloud protocols were employed to detect variability in scoring judgments of teachers at three levels of writing proficiency. Based on the results of Study One, twelve pre-service EFL teachers at three levels of writing proficiency were identified. The student scripts under real examination conditions were collected from one intact class of one key senior high school. The twelve participants were required to take approximately fifty minutes to rate twelve selected student scripts while at the same time conducting concurrent thinking-aloud protocols to report their thought processes during rating. The interview data were also collected to facilitate the interpretation of qualitative data.

Data analysis comprised two main stages. Firstly, MFRM was used to analyze the scores awarded by the participants. The transcribed think-aloud protocols were coded and frequency data of each category of the coding schemes were analyzed quantitatively as the basis for comparison across teachers at three levels of writing proficiency. Second, interview data were analyzed thematically to identify significant dimensions of differences or similarities across the three groups. The quantitative picture of scoring judgments of pre-service EFL teachers was further combined with qualitative details regarding their scoring

processes. Table 1-1 summarizes the research questions, data collection and analysis methods of the two studies.

Table 1-1 Overview of research design of the two studies

Research Questions	Data Collection	Data Analysis
<i>Study One: Writing Proficiency of Pre-Service EFL Teachers (N=81)</i>		
How does the writing proficiency of pre-service EFL teachers vary in terms of the discourse features inherent in their writing performance?	Writing performance: Writing performances of 81 participants in response to a designed writing test with three assessed levels Discourse features: 19 discourse features representing L2 writing construct were selected based on the document analysis	Examining differences of writing performances in terms of selected discourse features: • Descriptive statistics • Analysis of Variance
<i>Study Two: Scoring Judgments of Pre-service EFL Teachers (N=12)</i>		
How do pre-service EFL teachers at different levels of writing proficiency differ in their scoring judgments of student writing?	Quantitative scores Scores awarded by the participants Qualitative data • Think-aloud protocols of scoring the selected twelve scripts of secondary students • Immediate follow-up interviews with each of the twelve participants	Quantitative data analysis Descriptive statistics • Many-Facet Rasch Model • Kruskal-wallis H Test Qualitative data analysis • Thematic analysis