Foreign Language
Culture
Teaching

外语 文化 教学论丛

Exploring Rater Variability in Language Performance Assessment

语言运用考试中的评分员误差研究

张 洁◎著

by Jie ZHANG

ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

Exploring Rater Variability in Language Performance Assessment

语言运用考试中的评分员误差研究

张 洁◎著

by Jie ZHANG

ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

图书在版编目(CIP)数据

语言运用考试中的评分员误差研究 = Exploring Rater Variability in Language Performance Assessment: 英文 / 张洁著. —杭州：浙江大学出版社，2016.4

ISBN 978-7-308-15743-8

I. ①语… II. ①张… III. ①语言运用—考试—评分—误差—研究—英文 IV. ①H0

中国版本图书馆 CIP 数据核字(2016)第 067953 号

语言运用考试中的评分员误差研究
Exploring Rater Variability in Language Performance Assessment
张 洁 著

责任编辑 张颖琪
责任校对 陈丽勋
封面设计 项梦怡
出版发行 浙江大学出版社
(杭州天目山路 148 号 邮政编码 310007)
(网址: http://www.zjupress.com)
排 版 杭州中大图文设计有限公司
印 刷 杭州杭新印务有限公司
开 本 710mm×1000mm 1/16
印 张 11.75
字 数 283 千
版 印 次 2016 年 4 月第 1 版 2016 年 4 月第 1 次印刷
书 号 ISBN 978-7-308-15743-8
定 价 35.00 元

浙江大学出版社发行中心联系方式：0571-88925591；http://zjdxcbs.tmall.com

Foreign Language
Culture
Teaching

外语 文化 教学论丛

序　言

语言运用测试(language performance assessment)是一种对语言综合应用能力较为直接的测量方式。这类测试方式具有较高的真实性，且对教学有正面的促进作用，因此被广泛运用于众多大型考试中。然而这类测试任务通常需要评分员根据所制定的评分标准做出自己的主观判断，这对于考试的信效度和公平性都有很大的影响。

张洁博士的专著针对语言运用考试中的评分误差问题，主要探讨评分员误差对考试信效度的影响、误差的主要类型及造成误差的可能认知因素，从不同视角全面讨论了语言运用测试中的评分员误差问题。此书不仅系统地梳理了语言测试领域中评分误差的相关研究，介绍了该研究方向经常使用的定性定量研究方法，还通过两个实证研究翔实、具体地说明了如何运用不同的研究方法分析、测量和探究主观评分可能存在的误差及原因，在理论和实践层面都有较强的指导意义。书中包括的两个实证研究分别是她在硕士和博士阶段研究的主要内容，前者对应定量统计研究范式，运用多层面 Rasch 模型对四、六级口语考试的分数差异来源进行了系统研究;后者对应定性的以过程为导向的研究范式，对四级作文评分中评分员认知过程对评分准确度的影响进行了探讨。两个研究的主要发现对于大规模语言运用测试中评分误差的控制、测量以及更加有效地开展评分员培训和评分标准修订都具有十分重要的借鉴意义。

从 2004 年起涉足语言测试领域,张洁博士凭借其交叉的学科背景以及那份不可多得的灵气，已经成为中国语言测试研究的新生力量。该专著是她十余年辛勤耕耘的结果，更是她砥砺前行的动力。由衷地祝愿她的研究结出更为丰硕的成果。

何莲珍

2016 年 3 月

自　序

此书是我在硕士和博士两个阶段学位论文的基础上修改完成的。

这两个阶段的研究具有一定的延续性和互补性，前者运用统计方法对大规模口语考试中的评分员误差进行定量分析，后者深入探讨写作考试中影响评分员误差的内在思维过程因素。两个研究分别代表了评分员误差研究中两个不同的范式，较为全面地讨论了语言运用测试中的评分员误差问题。在浙江大学外国语言文化及国际交流学院攻读硕士学位和在广东外语外贸大学外国语言学与应用语言学专业攻读博士学位的五年时间，让我对应用语言学尤其是语言测试领域的理论构架、研究方法和前沿问题有了深入的了解和思考。两个阶段的研究尽管针对相似的主题，但采用了截然不同的研究方法。虽然本科期间的工科背景让我对于统计方法的运用驾轻就熟，但为了拓展自己的研究视野并熟练掌握各种研究方法，在博士研究阶段我选择以定性研究方法为主来更加深入地探讨评分员误差问题。实证阶段联系了十余位一线四级作文评分员以及阅卷点组长及专家进行有声思维和深度访谈，从中获取的千头万绪的文本材料曾一度让我感到理不清头绪。

在整个过程中，我要感谢我的硕士导师浙江大学何莲珍教授，在实证研究环节给予我极大的帮助和支持，并在我困惑和沮丧的时刻给我莫大的安慰与鼓励。还要感谢我的博士导师广东外语外贸大学曾用强教授，他对于博士生独到的直线思维训练让我能够在繁杂的文字材料中看清主线，明确研究问题并发现重要结论。也要感谢广东外语外贸大学的亓鲁霞教授，作为我博士论文的评审专家组成员，她给我的论文写作尤其是定性数据分析方面提出了许多宝贵意见。是他们的启示、引导、帮助和鼓励，让从我从半路转行的新手成为语言测试领域具有独立研究能力的研究者。最后，我还要感谢我的家人，是他们的陪伴与默默的支持让我能在学术研究的道路上走得更远，他们永远是我前进的动力和坚实的后盾。

2016 年 3 月于上海

Contents

Chapter 1　Introduction ······ 1

1.1　Rationales for studying rater variability ······ 1

1.2　Status quo of studies on rater variability ······ 2

1.3　An overview of this book ······ 5

1.4　Definition of key terms ······ 5

Chapter 2　Literature review: Studies on rater variability in language performance assessment ······ 7

2.1　Rater variability in language performance assessment ······ 7

2.2　Exploring rater variability using statistical analysis ······ 9

2.2.1　Introduction ······ 9

2.2.2　Rater reliability in Classical Test Theory ······ 9

2.2.3　Rater facet as variance component in Generalizability Theory ··· 10

2.2.4　Rater calibration in Many-Facet Rasch Model ······ 13

2.2.5　Summary ······ 21

2.3　Process-oriented approach to investigating rater variability ······ 22

2.3.1　Raters' decision-making: the "black box" behind the final ratings · 22

2.3.2　Indirect evidence ······ 23

2.3.3　Direct investigation of rating process: insights from verbal protocols ······ 27

2.4　Factors accounting for rater variability ······ 46

2.4.1　External factors ······ 47

2.4.2　Internal factors ······ 48

2.4.3　Situational factors ······ 52

2.5　A framework for comparison between rater groups ······ 53

2.6　Summary ······ 56

Chapter 3 **Study 1: Investigating the scoring reliability of CET-SET using Many-Facet Rasch Model** ······ 57

3.1 Issues in second language speaking assessment ······ 57

3.2 Challenges in test validation ······ 57

3.3 The context of the study ······ 63

3.4 Objectives of the study ······ 64

3.5 Methods ······ 65

3.5.1 Data ······ 65

3.5.2 Instrument (MFRM) ······ 66

3.6 Data analyses and findings ······ 67

3.6.1 Facet map ······ 67

3.6.2 Candidates ······ 69

3.6.3 Tasks ······ 74

3.6.4 Items ······ 75

3.6.5 Rating scales ······ 77

3.6.6 Raters ······ 80

3.6.7 Bias analysis ······ 82

3.7 Conclusions ······ 85

3.8 Implications ······ 87

3.9 Further research efforts to be made ······ 88

Chapter 4 **Study 2: Exploring how raters' cognitive and meta-cognitive strategies influence rating accuracy in essay scoring** ······ 90

4.1 Subjective scoring: A matter of reliability or validity? ······ 91

4.2 Exploring rating process: Looking into rater variability ······ 94

4.3 Rater cognition studies in writing assessment ······ 102

4.4 Methodology ······ 107

4.4.1 The context of the study ······ 107

4.4.2 Participants ······ 108

4.4.3 Materials ······ 108

4.4.4 Data collection ······ 109

4.4.5 Data analysis ······ 111

4.5 Results and discussion ······ 116

4.5.1 General patterns of differences in broad categories ··············· 116
4.5.2 In-depth investigation of differences in the major sub-categories ··············· 118
4.6 Summary and further discussion ··············· 130
4.7 Conclusion ··············· 134

Chapter 5 Conclusions ··············· 135
5.1 Summary of findings ··············· 135
5.2 Comparison of the two studies ··············· 136
5.3 Limitations ··············· 137
5.4 Further research efforts to be made ··············· 138

Appendix I CET-SET rating scale ··············· 140
Appendix II CET4 rating rubrics for the writing task ··············· 142
Appendix III The writing task of the Dec. 2006 administration of CET4 and range finders ··············· 144
Appendix IV Sample essays ··············· 147
Appendix V Instructions and training tasks for think-aloud session ··············· 149
Appendix VI Sample transcripts of raters' thinking aloud ··············· 151
Appendix VII Coding protocols for think-aloud verbal reports ··············· 154
Appendix VIII The coding scheme for raters' cognitive and meta-cognitive strategies ··············· 155

References ··············· 157
Index ··············· 172

List of Figures

Figure 2-1 Rater focus across proficiency levels ·· 31
Figure 2-2 A model of the decision-making process in composition marking ···· 44
Figure 2-3 A tentative model of factors affecting holistic scores of written composition (Sakyi, 2000) ·· 44
Figure 2-4 Model of factors affecting rating and score interpretation in oral assessment (Wang, 2007) ·· 45
Figure 2-5 A framework for comparison between raters (rater groups) ············ 54
Figure 3-1 An expanded model of oral test performance (Bachman, 2002a) ····· 62
Figure 3-2 Facet map ·· 68
Figure 3-3 Category probability curves for Item 1 ·································· 78
Figure 3-4 Category probability curves for Item 2 ·································· 79
Figure 3-5 Category probability curves for Item 3 ·································· 79
Figure 4-1 Characteristics of performance assessment (McNamara, 1996) ······· 93
Figure 4-2 An information-processing model of rating a composition (Freedman & Calfee, 1983) ··· 95
Figure 4-3 Variables affecting the rating process (Freedman & Calfee, 1983) ··· 98
Figure 4-4 Expanded model of scorer cognition (Wolfe & Feltovich, 1994) ···· 101

List of Tables

Table 2-1 Summary of general categories of textual features noted by raters ··· 29

Table 2-2 Descriptive framework of raters' decision-making behaviors when rating TOEFL writing tasks (Cumming et al., 2002) (final version, revised after coding new task and statistical analyses) ··· 34

Table 2-3 Differences in rating process between proficient and non-proficient raters ··· 41

Table 2-4 Model of rating procedure ··· 43

Table 3-1 CET-SET reported rank ··· 64

Table 3-2 Topic areas in November 2005 administration of CET-SET ··· 65

Table 3-3 Candidate report ··· 70

Table 3-4 Misfitting and overfitting candidates ··· 72

Table 3-5 Comparison between raw scores and adjusted scores ··· 73

Table 3-6 Task report ··· 74

Table 3-7 Item report ··· 76

Table 3-8 Rating scale calibration ··· 77

Table 3-9 Rater report ··· 81

Table 3-10 Summary of significant bias instances between raters and candidates ··· 83

Table 3-11 Bias report between raters and items ··· 84

Table 3-12 Bias report between raters and tasks ··· 85

Table 4-1 Background information of the participant raters ··· 108

Table 4-2 Score discrepancy distribution of expert ratings ··· 109

Table 4-3 MFRM statistics for ACCURATE and LESS ACCURATE rater groups ··· 113

Table 4-4 The major categories of the raters' rating behavior ··· 117

Table 4-5 The subcategories of the raters' error diagnosing strategies ··· 119

Table 4-6 The subcategories of raters' summarizing strategies ··· 124

Table 4-7 The subcategories of raters' monitoring strategies ··· 127

List of Abbreviations

ANOVA	Analysis of Variance
CAE	Certificate of Advanced English
CET	College English Test
CET-SET	College English Test-Spoken English Test
COE	Common European Framework
CPE	Certificate of Proficiency in English
CSW	Common Scale for Writing
CTT	Classical Test Theory
EFL	English as Foreign Language
EMT	English as Mother Tongue
ESL	English as Second Language
ESLPE	English as a Second Language Placement Examination
ESOL	English for Speakers of Other Languages
FCE	First Certificate in English
GT	Generalizability Theory
IELTS	International English Language Testing System
LTM	Long-Term Memory
MFRM	Many-Facet Rasch Model
NMET	National Matriculation English Test
NNS	Non-Native Speaker
NS	Native Speaker
OET	Occupational English Test
PETS	Public English Test System
SPEAK	Speaking Proficiency English Assessment Kit
STEP	Special Test of English Proficiency
TEM	Test for English Majors
TestDaF	Test of German as a Foreign Language
TOEFL	Test of English as a Foreign Language

TSE	Test of Spoken English
UCLA	University of California, Los Angles
WM	Working Memory

Chapter 1

Introduction

1.1 Rationales for studying rater variability

As a direct measure of learners' communicative language ability, performance assessment (typically writing and speaking assessment) is commonly espoused for its close link between the test situation and authentic language use and would often be taken for granted to enhance the validity of inference we could draw from the test scores (Bachman et al., 1995; Norris et al., 1998; Lynch & McNamara, 1998; Condon & McQueen, 2000; Bonk & Ockey, 2003). It has therefore been increasingly involved as compulsory or optional part in many large-scale language test batteries both at home and abroad (CET, PETS, NMET, TOEFL, IELTS, etc.).

However, the elicitation of complex response from examinees would inevitably call for human raters to make evaluative judgment on the effectiveness of test performance or the degree of mastery of the underlying construct the test sets out to measure. Research findings from numerous studies devoted to rater issue in the context of performance assessment have indicated that even after principled rater training or standardization, raters would still exhibit considerable variability or idiosyncrasies in the ratings they would award (Lunz et al., 1990; Lumley & McNamara, 1995; Wolfe, 1997; Weigle, 1998). Rater variability has therefore long been held as the most significant source of measurement error and potential threat to the reliability and fairness of performance assessment.

Furthermore, it is also well recognized that when engaged in the act of scoring, raters do not mechanically record what they see, rather, their ratings are rooted in observation, interpretation, and the exercise of personal and professional judgment (Myford & Wolfe, 2003). It would be reasonable to assume that, raters, with their internalized criteria and specific manner in implementing those criteria, would mediate between the test performance and the final score and determine, to a large extent, the meaningfulness of the score and the appropriateness of inference we could make from the test results. Rater variability, therefore, is not just a matter of reliability but holds a key position in determining the scoring validity of the whole test. As a result, detecting and measuring the degree of rater variability and exploring the underlying factors which would lead to the detected variation among raters are constantly regarded as one of the most important issues and a worthwhile focus of study in both research and practice in language performance assessment.

1.2 Status quo of studies on rater variability

There are two major orientations of studies on rater variability in the existing language testing literature. One is mainly concerned with how variability introduced by raters' subjective judgment would affect the scoring system (rater effect) through statistical analysis of the ratings awarded by different raters. The other is, on the other hand, more rater-oriented, the focus of which is therefore raters' rationales for their scoring decisions and the thought processes during their decision-making. Rather than focus on the final ratings, this approach to investigating rater variability would perceive raters as the decision makers who might follow different mental paths to arrive at their final judgment.

Of these two lines of research, studies focusing on statistical modeling of rater effect and investigation into the potential utility of different mathematic models have long been dominant in the literature. It might be that rater variability has been traditionally related with reliability issues and for many large-scale performance assessment the consistency and reliability of ratings is still the most practical and urgent concern. The most commonly utilized statistical techniques in detecting and measuring rater variability include inter-/intra-reliability indices in Classic Test Theory, estimation of variance component related with the whole rater facet in Generalizability Theory and calibration of individual raters' rating patterns in Many-Facet Rasch Model. By conceptualizing rater effect in different ways, these techniques provide different statistical indices depicting the quality of raters' ratings from different perspectives. Specifically, information could be obtained regarding how different raters would agree with each other in terms of their rank-ordering of the same group of examinees using inter-rater reliability coefficient (Standsfiled & Ross, 1988; Huot, 1990), the extent to which rater group as a whole would contribute to the total variance in the final scores using G-study (Bachman et al., 1995; Lynch & McNamara, 1998; Clauser et al., 1999), and how individual raters would differ from each other in their overall severity, self-consistency and significant bias towards examinees, tasks or items using calibrations from MFRM (Engelhard, 1994; Bachman et al., 1995; Lynch & McNamara, 1998; Condon & McQueen, 2000; Myford & Wolfe, 2000; Bonk & Ockey, 2003; Eckes, 2005).

Although the emergence of more sophisticated statistical methods like GT and MFRM enables the researchers to investigate rater effect in a more finely-tuned fashion than the traditional simple correlation coefficients for inter-rater reliability in CTT, statistics, in its very nature, is to break down the observable score variance into a single or limited number of dimensions, leaving the complexity and richness implicated in rating process unexplored, not to mention the various potential factors which might influence raters' rating behaviors. Therefore, statistical modeling would always leave some sources of score variance unexplained, which researchers could hardly interpret but label them as "idiosyncrasies" of raters. That is why many researchers have called for more in-depth investigation into this "mysterious" and sometimes troublesome area of "idiosyncrasies" at the end of their quantitative studies (Douglas, 1994; Weigle, 1998; Eckes, 2005).

The other line of research, therefore, is devoted to exploring how raters would arrive at their final decisions, with the aim to find out the underlying reasons leading to the persistent idiosyncrasies among raters. Some of these studies draw upon indirect evidence to infer what performance features raters might attend to for their judgment, such as the correlation between various textual features and raters' ratings (Homburg, 1984; Ferris, 1994; Laufer & Nation, 1995; Cumming et al., 2006) and raters' comments or annotations on the target essays (Turner & Upshur, 1996, 2002; Jenkins & Parra's, 2003; Eckes, 2008). These endeavors help to extract important features in examinees' performance which might influence raters' decision-making and therefore provide useful information for the validation or development of rater-oriented rating scales. There are also other studies which utilize raters' concurrent or intro-/retrospective verbal protocols as direct evidence of their thought processes in their decision-making. Some of these studies mainly focus on describing the similarities and differences in raters' text focus (heeded information during rating) and their reading or rating styles as well as the strategies and behavior conducted to acquire and process the heeded information (Vaughan, 1991; Milanovic et al., 1996; Deremer, 1998; Sakyi, 2000; Cumming et al., 2001, 2002; Lumley, 2002, 2005; Orr, 2002; Brown et al., 2005; Hubbard et al., 2006; Wang, 2007). Nevertheless, given that the nature of such studies is exploratory and descriptive, their findings are mixed due to specific assessment contexts and different rater groups investigated. This is quite natural in that, as a complex cognitive process and ill-structured problem-solving task, rating is bound to exhibit considerable variation across different raters and conditions. However, what really matters is not just describing the superficial similarity or difference observed in raters' behaviors, but to investigate what would contribute to make them behave in that way.

Though small in number, there emerged some studies which began to investigate how raters' personal characteristics such as experience and expertise in assessing writing (Huot, 1993; Pula & Huot, 1993; Wolfe, 1997; Wolfe & Feltovich, 1994; Wolfe & Ranney, 1996), L1 status (Erdosy, 2004), professional background (Cumming, 1990; Cumming et al., 2001, 2002; Erdosy, 2004), and knowledge of examinees (Deremer, 1998) would influence raters' decision-making by comparing different groups of raters. Compared with studies mainly concerned with describing salient patterns of rating process of different raters, comparing groups of raters, as mentioned above, is a step further in rater cognition study, which have introduced an element of experimental design into the study to explore the underlying factors accounting for the detected differences in raters' decision-making. However, there are hardly any attempts, except Wolfe et al.'s studies (Wolfe, 1997; Wolfe & Feltovich, 1994; Wolfe & Ranney, 1996), to link rater variability in their decision-making process with their actual rating performance (i.e. how accurate and consistent their ratings would be compared with some pre-defined norm), which would be of primary concern in large-scale standardized performance test settings. Furthermore, these studies, although having revealed important dimensions of difference among raters with different personal characteristics, still fall short to provide a unified account for the mechanism of how these factors could induce raters' variability in their decision-making. Without such knowledge, we are likely to be overwhelmed with the diverse and sometimes inconsistent findings derived from different contexts. It is therefore necessary to probe further into raters' mental basis for decision-making and how the variability inherent in their minds would lead to the variability in their decision-making process and outcome, thereby advancing our