CET4

作文评分人混合型反馈的效果研究

徐鹰 著

Investigating the Effectiveness of Mixed Feedback to CET4 Essay Rater Performance

华南理工大学出版社
SOUTH CHINA UNIVERSITY OF TECHNOLOGY PRESS

“2014年广州市哲学社会科学‘十二五’规划课题”（批准号 14Q11）研究成果

CET 4
作文评分人混合型反馈的效果研究

徐鹰 著

华南理工大学出版社
SOUTH CHINA UNIVERSITY OF TECHNOLOGY PRESS
·广州·

图书在版编目（CIP）数据

CET4 作文评分人混合型反馈的效果研究/徐鹰著. —广州：华南理工大学出版社，2014.8
ISBN 978-7-5623-4360-8

Ⅰ.①C… Ⅱ.①徐… Ⅲ.①大学英语水平考试-写作-评分-研究 Ⅳ.①H315-42

中国版本图书馆 CIP 数据核字（2014）第 184196 号

CET4 作文评分人混合型反馈的效果研究
徐鹰 著

出 版 人：韩中伟
出版发行：华南理工大学出版社
（广州五山华南理工大学 17 号楼，邮编 510640）
http://www.scutpress.com.cn E-mail：scutc13@scut.edu.cn
营销部电话：020-87113487 87110964 22236378 87111048（传真）
责任编辑：朱彩翩
印 刷 者：虎彩印艺股份有限公司
开 本：787mm×1092mm 1/16 **印张**：15 **字数**：377 千
版 次：2014 年 8 月第 1 版 2014 年 8 月第 1 次印刷
定 价：40.00 元

序 言

本专著是徐鹰博士在博士学位论文的基础上修改而成的。

徐鹰博士的学位论文采用反向平衡实验设计，在连续两次大学英语四级考试（CET4）作文评分周期内（间隔为半年）对评分人混合型反馈的有效性进行系统研究，旨在探讨评分人混合型反馈能否帮助评分人获得评分技能的理论和实践问题。评分人反馈是评分人培训中常用的一种方法，是保证考试信、效度的重要手段。目前国际语言测试界围绕这一课题开展了一些实证研究，但大多数研究实验设计不够严谨，只提供一次性反馈，仅关注作为评分终端产品的分数，忽略了评分人决策行为的变化，故影响了评分人反馈的效果和研究质量。国内语言测试界在该领域的研究尚属空白。徐鹰博士提出的心理测量学和阐释学双重研究视角别具新意。评分人混合型反馈不仅包括对分数的多层面 Rasch 模型分析结果，还增加了专家评分人的给分理据。此外，他还采用了以有声思维为代表的质性研究手段，力求建立中国特定考试评分背景下评分人培训和反馈的新型模式。因此这项研究不仅具有明显的理论探索意义，而且具有广泛的实际应用价值。徐鹰博士的学位论文选题起点较高，相关文献翔实，实验方法、实验设计、实验材料、实验工具、实验步骤均符合学术规范，实验结果统计分析、定性数据解读具有相当高的信度，对研究结果也作出了符合逻辑的讨论和理论阐述。论文行文流畅，结构符合规范。总体而言，该研究通过对 CET4 作文评分人提供系统反馈以控制评分误差，这一尝试具有填补空白的意义，对我国乃至国际大规模考试具有重要启示。

徐鹰博士能在规定学习期限内完成学业实属不易，这与他对语言测试专业方向的浓厚兴趣和善于利用时间的学习方式密不可分。在三年博士学习期间，他积极进取，追踪学术热点，培养问题意识，不断在教学和科研实践中挖掘新的学术兴趣点；在潜心读书的同时，他还笔耕不辍，在《天津外国语大学学报》《中国考试》《外语测试与教学》《华南理工大学学报（社会科学版）》等知名期刊上发表了若干篇论文，同时成功申请了广州市社科规划青年项目和广东省教育科学“十二五”规划项目各一项。作为他的博士导师，我对他取得的上述成绩深感欣慰，并希望他毕业后能从零开始，随时警惕浮躁心情和急功近利思潮，坚定地沿着自己的研究方向大步前进。持之以恒，必有所获。

愿徐鹰博士今后取得更大的学术成就！

是为序。

曾用强

2014 年 6 月

前 言

评分人差异是威胁做事测试分数信、效度的主要来源。人们通常提供评分人反馈以求控制这种测量误差。过往文献中针对评分人反馈效果的研究得出的结论不一，揭示出人们对评分人反馈所赋予的积极期望和其有限作用之间的巨大落差。该现象主要原因在于：首先，评分人反馈的内容局限于分数的定量分析结果；其次，评分人通常得到一次性反馈，缺乏系统性。迄今为止，人们对提供多次的混合型反馈（结合分数定量分析结果和标准分数的评分理由）有效性缺乏认识。为填补这一研究空白，本研究在心理测量学和解释学双重视角观照下安排评分人反馈内容，具体包括分数的多层面 Rasch 模型分析结果和专家评分人提供并经过验证的评分理由。同时，在一个为期九天的大学英语四级（CET4）作文评分周期内，连续提供三次该混合型反馈。

本研究采用了“前测 + 后测”的准实验设计。由于客观条件所限，实验组和对照组评分人没有实现随机分组。为了控制可能产生的顺序效应，本研究同时采用了反向平衡设计。研究参与者主要包括 25 位通过 CET4 作文评分培训考核的评分人。他们被分为 A 组（13 人）和 B 组（12 人）。在第一个 CET4 作文评分周期内，A 组收到了混合型反馈，B 组没有收到该反馈；在第二个 CET4 作文评分周期内，A、B 组角色对调。本研究数据主要包括评分人给分、评分人对自己给分的把握度、部分评分人完成的有声思维、实验组评分人完成的两份问卷（反馈内容理解问卷和反馈效果评估问卷）以及某些实验组评分人所做的后续访谈。数据分析主要包括在小组和个体层面上所做的定量分析和/或定性分析。评分人给分用多层面 Rasch 模型进行分析，并以专家评分人给分作为定锚。评分人的给分把握度用 ANOVA 进行组间比较，并用 t 检验进行个体对比。评分人的有声思维经过转写、切分后按照文本特征和信息处理行为进行编码。四种构念相关特征所占比例和四种主要信息处理行为所占比例分别用 MANOVA 进行组间比较。两份问卷的结果进行描述性统计分析。访谈结果进行主题分析。

本研究发现：首先，在两次评分周期内，这种混合型反馈未能有效减少评分人差异，具体体现为严厉度、内在一致性、对考生的偏差、趋中倾向以及严厉度漂移；其次，它未能有效调整评分人的严厉度和偏差，尽管在第一个评分周期内它有效调整了评分人的内在一致性；再次，它未能影响评分人的给分把握度。但是它能有效提高评分人有声思维中构念相关特征（尤其是连贯）比例和这些特征的种类，从而提高分数的效度；最后，绝大部分评分人认为这种混合型反馈内容可以理解，同自己的真实表现一致并且可以直接运用，因此对反馈作用表示肯定。该混合型反馈未能产生作用的原因从评分任务要求（如工作量要求、工作时间要求和评分人绩效评估制度）、评分标准特点、模拟作文特点和评分人特点等方面进行了探讨；该反馈对评分人所关注的文本特征产生作用的原因从反馈内容和反馈方式的改进、CET4 作文评分人对反馈的内在需求以及 CET4 作文评分标准的特点等方面进行了分析。通过对上述结果建立内在联系，我们发现不仅评分人表现受到各种评分任务要求的影响，而且评分人特征

和评分人培训（包括反馈）相互影响，但是这些特点在 Knoch（2009）的扩展型做事测试模型中未能体现。

本研究结果说明，虽然这种混合型反馈未能减少评分人差异，但能提高分数的解释力，从而提高分数的效度。

本研究是迄今为止国内语言测试界对大规模考试评分人反馈进行研究的第一本专著，具有一定的理论价值和实践意义。但由于作者水平有限，书中错漏之处在所难免，敬请广大读者批评指正。

徐鹰
2014 年 6 月

Preface

Rater variability poses a serious threat to score reliability and validity in performance assessments. Rater feedback is generally viewed as a useful measure to guard against this measurement error. Research on the effectiveness of feedback has produced mixed findings, which reveals a mismatch between the positive expectation held for rater feedback and its limited effects. Reasons are mainly two-fold: first, feedback content is dominated by quantitative results of ratings; second, feedback is always provided in a snapshot manner. To date, little is known about the effectiveness of mixed feedback combining the psychometric analysis represented by the many - facet Rasch model (MFRM) result of raters' ratings and the hermeneutic comment on validity scores over several iterations in a rating session lasting a few days. To fill the research gap, the present study prepared the rater feedback by enriching the MFRM analysis of ratings with the expert rater's validated rationales, and delivered such feedback repeatedly over three iterations in a whole CET4 essay rating session lasting nine days.

A quasi - experimental design with a "pretest-posttest control-group design" was adopted. As random sampling of raters was unrealized due to practical constraints, a counter-balanced design was used in order to control the order effect. 25 CET4 accredited raters were divided unevenly into Group A (13 raters) and Group B (12 raters). In the first rating session, Group A received the mixed feedback and Group B received no feedback. The roles of Group A and Group B in the second rating session were reversed. Five data sources were collected including ratings, raters' degree of certainty for each rating, some raters' think-aloud protocols (TAPs), the experiment group raters' answers to the questionnaire of feedback perceptions and the questionnaire of feedback effectiveness evaluation, and a follow-up interview with some experiment group raters. Data were analyzed either quantitatively or qualitatively at both group level and individual level. Ratings were analyzed using MFRM anchored with the expert rater's ratings. Raters' degree of certainty were analyzed using a 2×2 ANOVA at the group level and a t-test at individual level. TAPs were coded within the coding scheme of text features and the coding scheme of information-processing behaviors. Percentages of four construct-related features and four major information-processing behaviors were analyzed using MANOVA. Results of the two questionnaires were analyzed descriptively and interviews were analyzed thematically.

It was found that at the two rating sessions, the mixed feedback was ineffective in decreasing rater variability in terms of severity, consistency, bias, central tendency and severity DRIFT (differential rater functioning over time). It was ineffective on raters' successful adjustment to severity and bias, although its effectiveness on raters' successful adjustment to consistency was half validated in the first rating session. It was ineffective on raters' degree of certainty. It was effective in improving raters' mean percentage of construct-related features, esp. the percentage of coherence, and also in helping raters adopt a wider coverage and a heavier weight of construct-related features in making scoring decisions. The vast majority of raters deemed that the mixed feedback was understandable, consistent with their real performance and able to be acted upon, and thus held a

positive attitude towards its effects. Reasons why the mixed feedback was ineffective on raters' variability, successful adjustment to severity and bias, and degree of certainty were discussed in terms of the task requirements (the workload, the duration of the rating session and the performance appraisal system), the nature of rating scale, and the characteristics of mock essays and the characteristics of raters. Reasons why it was effective on raters' heeded text features were discussed in terms of its content and delivery mode, the internal needs of CET4 essay raters in rating practice, and the nature of CET4 essay rating scale. Finally, an internal link was established between findings for different research questions, which uncovered the truth that not only a number of task requirements would influence raters' performance, but also there is an interaction between rater characteristics and training (including feedback). However, they are not explicitly manifested in Knoch's (2009) expanded model of performance assessment.

Findings from this study suggest that although rater variability wasn't diminished after feedback, score validity was improved. Therefore, interpretability of scores can be reinforced and a stronger validity argument for scores can be formulated.

This book makes the first step to investigate rater feedback in the large scale language tests at home. It is both theoretically and practically significant to some extent. I hope to have had all the errors eliminated; the responsibility for those that may be left is entirely mine.

Xu Ying
June, 2014

Acknowledgements

There are many people to whom I need to express my gratitude, without their help this study would be an impossible mission.

First, I feel deeply indebted to Prof. Zeng Yongqiang, my supervisor at Guangdong University of Foreign Studies, who made my doctoral study possible, inspired me throughout the whole process and supported me whenever I needed his help. His penetrating insights at lectures and seminars have aroused my curiosity and ignited my passion to inquire scientifically into the truth. Needless to say, his role as "the guide on the side" would be a lighthouse for my academic odyssey. Besides, it should be mentioned that without Prof. Qin Xiubai's recommendation and Associate Prof. Zhang Fengchun's introduction to him, I would not have the life-changing chance.

I am also thankful to my research committee members, Prof. Liu Jianda, Prof. Li Qinghua and Doctor Cai Hongwen, for their constructive and valuable advices which undoubtedly have improved the overall quality of the dissertation.

I owe a lot to Prof. Qi Luxia, without whose advices in the design stage, this study would never be completed so smoothly.

I must express my heartfelt gratitude to CET Guangzhou Marking Center and its director, Associate Prof. Zhang Huaijian for providing opportunities and facilities that enabled me to collect data without any problem.

I am grateful to all the participants in the study, particularly the CET4 essay raters. They have remained anonymous, but for me they are genuine friends who contributed their time and energy, and thus worthy of my respect.

I would like to thank other PhD students at Guangdong University of Foreign Studies, especially Wang Weiqiang, Zhou Yan and Xu Liu for their valuable advices and kind-hearted help in the analysis of data.

Finally, my deepest gratitude goes to my wife, my little daughter, my mom and my parents-in-law. Their support and encouragement throughout my study have been my source of strength over the past years. My daughter has grown up with little care from me, and I owe her more than I can possibly give back.

Contents

Contents

Contents

Contents

Chapter 1 Introduction

1.1 Background of the study

The field of language testing and assessment has switched its attention to performance assessments since the "communicative turn" in the early 1980s (McNamara, 1996; Bachman, 2000). As the key factor in the performance assessment, raters are supposed to provide appropriate ratings and do so in a consistent way (Lim, 2011). The implication is two-fold: on the one hand, raters should provide reliable scores which are accurate, reproducible and generalizable to other testing occasions and other similar test instruments (Ebel & Frisbie, 1991); on the other hand, raters should carry out scoring in a manner consistent with the construct and measurement goals in order to support a validity argument for scores (Bejar, 2012).

However, the perennial problem of rater variability is a major threat to test reliability as well as test validity of the inferences drawn from the testing results. Hence, rater effects have become an integral part of performance assessments (Chalhoub-Deville, 1995; Upshur & Turner, 1999). It can manifest itself in a number of ways, from leniency/severity, central tendency, randomness, halo effect to differential leniency/severity (Myford & Wolfe, 2003). In short, the error-prone nature of rater facet leads to unreliable, invalid and unfair results. In the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999), Standard 1.8 documents that: "When statements about the process employed by observers or scorers are part of the argument for validity, similar information should be provided" (p. 19). Since scoring is at the core of both interpretative argument and validity argument (Myford, 2012; Suto, 2012; Bejar, 2012), therefore, it is necessary to measure, detect and control rater variability, particularly in the large-scale test.

Rater training is a popular method, which is often regarded as a crucial component because it is believed to compensate for different examiner backgrounds, and adjust examiner expectations so that any variability in the marking process caused by divergent expectations is diminished (Charney, 1984; Huot, 1990). Rater training is also defined as the process where judges are (re-) introduced to the assessment criteria and then required to rate a number of writing samples according to these criteria in an attempt to arrive at a common interpretation of their meaning (Elder, Knoch, Barkhuizen, & von Randow, 2005). Generally speaking, training is often used as a means to familiarize raters with the tasks and the rating criteria. The most pronounced findings of training effects are based on Weigle (1994, 1998), who discovered that rater training had the effects of clarifying the intended rating criteria, modifying rater expectation and heightening concern for inter-rater agreement. Furthermore, it was more successful in helping raters award more predictable scores than in getting them to give identical scores. However, there has been little empirical research to assess which elements of a training program are effective and why (Fulcher & Davidson, 2012,

p. 417). Empirical studies of rater training are far from exhausted. Hence, as Weigle (1998) indicated, little was known about what actually occurred during examiner training and how it affected the examiners themselves.

To be specific, rater training is usually made up of familiarization activities, practice rating, feedback and discussion (Lane & Stone, 2006). Feedback is generally regarded as part of rater training, which may impact on rater consistency. With the emergence of Item Response Theory (IRT), particularly the introduction of many-facet Rasch model (MFRM), the individualized feedback to raters is made possible. A number of empirical studies (Stahl & Lunz, 1991; Wigglesworth, 1993; Lunt, Morton, & Wigglesworth, 1994; Hoskens & Wilson, 2001; Elder et al., 2005; O'Sullivan & Rignall, 2007; Knoch, 2011) have been done in an attempt to enhance their rating accuracy. However, due to different methodological design and research contexts, these studies engendered mixed findings. A critical review of these inquiries revealed that they emphasized too much on the diagnostic function of feedback, yet ignored its hermeneutic side. This defect in research design has a detrimental effect in two aspects: first, when it comes to research objectives, changes in the rating product (represented by the scores) are emphasized, while changes in the rating process (how raters arrive at the scores) are passed over. It is highly possible that providing feedback may not bring a change to the MFRM statistics generated out of the scores, but may make a difference to raters' decision-making process. Second, in terms of feedback content, the effectiveness of feedback may be deteriorated to a large extent by providing IRT-based results but without any explanation on how the validity scores are reached. The reason is that raters can become well informed of their rating pattern, but they have no way to improve as a result of lacking explanation why they under-perform in the field, not to mention how to improve their rating expertise. Therefore, raters have to return to their own way of rating in practice. In passing, all of the above empirical studies of rater feedback were carried out in the oversea context. To my knowledge, no study to date on the effectiveness of feedback to raters for any tests in China was reported. As a consequence, since there is no evidence of the effectiveness of mixed feedback with a combination of MFRM results and the expert's explanation for the validity scores, the present study aims to fill the research gap.

1.2 Purpose of the study and research questions

The general purpose of the study is to inquire into the effectiveness of mixed feedback combining MFRM analysis of ratings and the expert rater's validated rationales for his/her ratings in the real rating context of College English Test Band 4 (CET4) in China with a quasi-experimental design (Shadish, Cook, & Campbell, 2002; Dörnyei, 2007). Due to the restriction on access to the test data, this research would have to use mock essays rated by two groups of accredited CET4 raters. Mock essays are written by students who just attended the then administration of CET4. The whole rating process would be embedded within the genuine rating context of CET4 writing. In other words,

except essays, other elements in the real context, raters and the time of rating to name a few, would be kept all the same to those in the real CET4 essay rating context. Therefore, it can be seen as a replication of the real rating process. This research design is basically quasi-experimental because it is hardly possible to use random sampling due to practical constraints.

To guide the research, the following five questions are raised:

Under the circumstances where the mixed feedback to rater performance is provided over 3 successive iterations in a 9-day CET4 essay rating session,

1. What effect does the mixed feedback have on rater variability in terms of
   (1) severity?
   (2) internal consistency?
   (3) bias against students?
   (4) central tendency?
   (5) severity differential rater functioning over time (DRIFT)?
2. To what extent can the experiment group (EG) raters incorporate the mixed feedback into rating behavior when it is provided repeatedly over 3 iterations?
3. What effect does the mixed feedback have on raters' degree of certainty for their ratings?
4. What effect does the mixed feedback have on raters' decision-making in terms of
   (1) text features?
   (2) information processing behaviors?
5. What are raters' perceptions of such feedback?

1.3 Significance of the study

The study is meaningful both in its contribution to a better understanding of the nature of scoring expertise and the usefulness of mixed feedback, and in its potential use in real practice, esp. in the training, monitoring and evaluating raters in the performance assessment.

1.3.1 Theoretical significance

As providing feedback to rater performance over a whole rating session lasting a few days (particularly in the essay rating context of China) is never attempted, hence, the present study adds to our knowledge by looking into the key issues of acquisition and maintenance of the scoring expertise with the help of the mixed feedback. In a word, it has theoretical value by answering the question: whether the scoring expertise can be gained and maintained by the mixed feedback? This is the de facto nature of teachability and learnability of scoring expertise.

To be more specific, the major theoretical value is two-fold. First, the study is epistemologically beneficial to further our understanding of rater feedback effectiveness and the nature of scoring expertise. Since there is little research systematically investigating into the effects of training and its

elements (feedback in particular), it is less clear whether detailed feedback to rating performance can enhance rating quality (Fulcher & Davidson, 2012, p. 417), let alone raters' decision-making behaviors and their psychological traits such as degree of certainty. Hence, the present study would try to verify the effectiveness of feedback to rater performance as a method in rater training by adopting a systematically designed quasi-experiment. It tries to gain a comprehensive view on the effectiveness of the mixed feedback and inspect the interrelationship between raters' change in scores, decision-making behaviors and degree of certainty at both group level and individual level, in an attempt to establish an intrinsic connection. Although random sampling is unachievable in practice, a counter-balanced design would be adopted in order to establish a cause-and-effect relationship. Second, since the testing circle acknowledges that raters fall into fixed rater types in the interpretation and use of routinely-based rating rubrics (McNamara & Adams, 1991; McNamara, 1996; Milanovic, Saville, & Shuhong, 1996; Sakyi, 2000; Cumming, Kantor, & Powers, 2002; Eckes, 2008), hence whether raters' behavior can be changed via feedback becomes the core issue. Regrettably, studies along this line are underdeveloped. Therefore, findings in the present study are meaningful to verify or falsify rater type hypothesis as such.

1.3.2 Methodological significance

In addition, the study is methodologically enlightening by enriching rater feedback with a combination of statistical analysis and the expert rater's rationales for validity essays (referred to essays whose scores have been validated and can be used as the norm). As the traditional way of rater training, characterized by the dependence on the Classical Test Theory (CTT), is vehemently criticized as damaging validity of the ratings by inducing raters to attend to superficial features of examinee performance (Shohamy, 1995; Reed & Cohen, 2001; Hamp-Lyons, 2007), the mixed feedback in this study is both practically feasible and theoretically valid because its implementation is manageable in the real CET4 essay rating practice, and its content is backed by both psychometric and hermeneutic evidence.

As the present study employs a mixed-methods research design, it makes the fullest use of qualitative data and quantitative data to contribute to a deeper understanding of effectiveness of the mixed feedback. As the problems concerning raters are complex and the use of either quantitative or qualitative approach alone is inadequate to address this complexity, hence the mixed-methods approach is adopted which is expected to provide an in-depth understanding of research questions.

1.3.3 Practical significance

Finally, its significance in practice roots in the usefulness of such mixed feedback. There is no study on the effectiveness of rater feedback under the condition of extraordinarily heavy workload and over a succession of several days. However, the rating conditions in most tests in China are characterized by the heavy workload. For example, in the real CET4 essay rating practice, a rater has