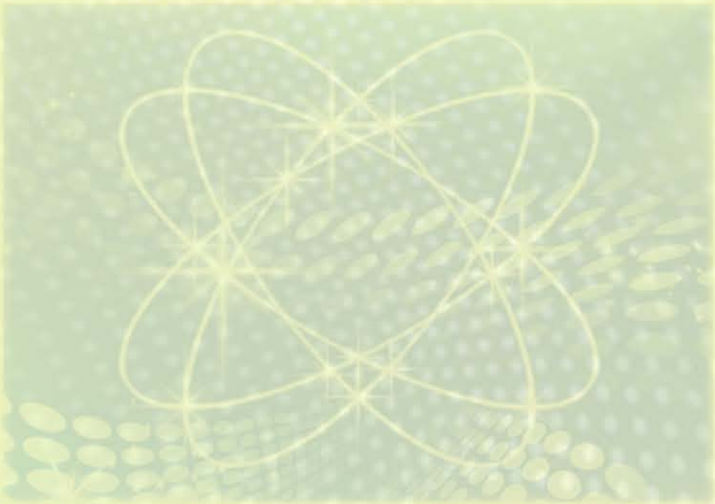
基于Rasch模型的英语阅卷人阅卷行为差异研究

尹晓腾　著



天津科学技术出版社

基于Rasch模型的英语阅卷人阅卷行为差异研究

尹晓腾 著

天津出版传媒集团
天津科学技术出版社

图书在版编目（CIP）数据

基于Rasch模型的英语阅卷人阅卷行为差异研究：英文 / 尹晓腾著. —天津：天津科学技术出版社，2018.4

ISBN 978-7-5576-5022-3

Ⅰ. ①基… Ⅱ. ①尹… Ⅲ. ①英语—考试—标准—研研—英文 Ⅳ. ①H319

中国版本图书馆CIP数据核字（2018）第073930号

责任编辑：布亚楠

天津出版传媒集团

天津科学技术出版社出版

出版人：蔡 颢

天津市西康路35号 邮编300051

电话（022）23332695（编辑部）

网址：www.tjkjcbs.com.cn

新华书店经销

天津午阳印刷有限公司印刷

开本787×1092 1/16 印张14.75 字数410 000

2018年4月第1版第1次印刷

定价：32.00元

ABSTRACT

In a writing performance assessment, raters' roles are one of themost crucial factors for fair assessment. Although more research studies of decision-making behavior are desperately needed, stake-holders including administrators, teachers, parents, and students have been more concerned about assessment scores rather than about raters' scoring behavior. Recently, most Chinese universities have both native English speaker (NES) and non native English speaker (NNES) teachers to teach and assess English writing skills.

With the very context in mind, this study aimed to investigate the scoring behaviors of NES and NNES raters in English writing assessment. For this research, 14 raters (7 NES and 7 NNES raters) were recruited to score 39 essays written by university students on a single topic selected from a list of topics available in TOEFL (Test of English as a Foreign language) preparation materials. The raters were asked to assess the essays using both analytic and holistic rating scales. In addition, raters were also required to write more than three comments explaining why they assigned those scores to the essays, using the holistic assessment scale.

In order to analyze the raters' assessment behaviors, this study examined scoring differences between the two groups in terms of severity, consistency, and bias, using a Many-facet Rasch measurement (MFRM) analysis. Based on grounded theory, the raters' written comments were categorized to understand their latent perspectives on the writing assessment. The differences in the frequencies of categories were examined through a chi-square test for any statistical significance.

The findings were that most NES and NNES raters tended to score leniently in both analytic and holistic scoring procedures. According to the Rasch modeling analysis of the holistic score data from each rater group, NESs were more severe raters than NNES s, NNES being more lenient raters than NESs. Both groups of NES and NNES raters had high

levels of inter-rater consistency in the two scoring assessment procedures. NNES raters had low levels of intra-rater consistency, while NES raters' intra-rater consistency in both scoring procedures was high. The Spearman-Brown rank correlations among the holistic score variable and the five analytic score variables based on the five criteria provided in a rubric (i. e., *Contents, Organization, Grammar, Wordchoice*, and *Mechanics*) were conducted. There sult of this correlation al analysis for each group indicated that the variable of *Contents* had the highest correlation with the holistic score variable in the NES rater group. However, the highest correlations between the holistic score variable and *Contents* and *Organization* were observed in the NNES rater group. The data analysis of the written comments revealed that NES and NNES raters approached their decision-making with different points of view on the writing assessment. NNES raters gave more weights to Organization and General than NES raters, but NES raters to the other categories of Contents, Grammar, Vocabulary, Mechanics, and others.

In this study, it was found that NES and NNES raters assigned scores to essays differently, producing some variation in students' scores due to different perspectives. In addition, it was confirmed that NES raters own assessment criteria categorized from their written comments exactly corresponded with the analytic criteria employed in this study. Some implications for improving the assessment of essays are suggested including the necessities for rater training and the potential benefits of using the analytic scoring procedure over the holistic scoring procedure.

Keywords: writing assessment, NES and NNES raters, assessment behavior, grounded theory, many-facet Rasch measurement.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

1. 1 Significance of the study

In recent years, students' English writing skills have been considered important in our profession of teaching English in the EFL (English as a Foreign language) context. People around the world have been writing each other through the internet, e-mail, SNS (Social Network Service), and other means of communication for sharing their ideas in politics, education, business, and others. Thus, in Korea, secondary schools, universities, and other educational institutions teaching English are attempting to improve their students' writing abilities.

However, China is some what unique in its approaches to teaching writing skills to students by putting too much emphasis on testing and assessment of writing skills rather than on some systematic development of students' writing skills. For instance, summative classroom test s are important for students from elementary school to university since test result s are always used for making decisions of many sorts for accepting or rejecting college applicants and hiring employees at a company, and so on. Because of the importance of testing and assessment, writing assessment in China has become crucial, whether writing ability is assess ed in indirect methods (e. g. filing in blanks with words or expressions) or direct methods (e. g. writing an essay).

In China, most undergraduate students have to take an English writing course. Some are taught by non native Chi NES e teachers of English, and others by native English speaker teachers. The students at the same university may take the same " English writing " course taught by either non native Chi NES e teachers or native English speaker teachers. In this kind of situation, the teachers, regardless of their first language s, use the same rubric similar to that of the TOEFL (Test of English as a Foreign language) writing assessment. Yet, a

student may get a different score of the same writing sample, depending on the teacher's first language. Taking this variable into account, getting a final grade of 'A' or 'C' can make a big difference to those students and potentially in finding a job after their graduation from their colleges.

To add more complexity to the writing assessment, the students may get the same scores with a high level of test reliability from two groups of the native English speaker (NES) raters and non native English speaker (NNES) raters. However, the two groups of raters may have some different perspectives. For example, one essay can be awarded a point of 7 out of 10 on the Likert-type scale used by both NES and NNES raters but for different reasons: the NES raters may make positive comments such as 'well presented ideas', while the NNES raters may state that there have been 'many simple grammar mistakes' (Shi, 2001). Concerning this type of variance due to different perspectives on assessing written products, Shi (2001) brought forth the importance of construct validity in writing assessment. researchers have examined the raters' perspectives when scoring students' written products. Some researchers studied experienced raters (Cumming, 1990; Cumming, Kantor, & Powers, 2002; Eckes, 2005; Lumley, 2002; Wiseman, 2012), while others English as a Foreign language (EFL) raters (Barkaoui, 2007). Still, there are relatively few research studies that have examined rater scoring behavior or decision-making process by NES and NNES raters. furthermore, there has been a dearth of research on direct comparisons of rater scoring behavior between NES and NNES raters. Although Shi (2001) examined the rating behavior of NES and NNES raters in writing assessment, there search was conducted through using the holistic scoring method with a10-pointscaleandanalysed with Multivariate analysis of Variance (MANOVA) to investigate the mean differences between the two rater groups on the ratings. However, there search failed to use proper statistical analyses for measuring some differences in rating behavior.

First, the present study applied Many-facet Rasch measurement (MFRM) to understand rating tendencies of two groups of NES and NNES. MFRM is available in FACETS, the statistical software program fit to investigate some influences of various variables on test

scores, while MANOVA shows simply score differences in two groups.

For example, the FACETS program can explore the differences in raters' severity, consistency, and even bias concerning other facets such as students' writing ability, criteria difficult y, and task difficulty.

The FACETS program in this study allowed there searcher to investigate scoring behavior in more detail.

This present study also examined whether or not scoring of essays based on the analytic scoring criteria corresponded with written comments concerning why certain scores were assigned to each paper.

While the amount of correspondence between analytic scoring criteria and written comments may indicate a high level of construct validity, few studies have been conducted on this theme. Thus, in an effort to analyze rating behavior based on written comments made by the raters, each comment was classified and sub classified in search of meaningful token patterns. Then, patterns were compared with the analytic criteria in this study. In addition, the correlations between the holistic scores and scores of each criterion in the analytic scoring method were explored to understand which criterion in the analytic scoring method had the greatest impact on the holistic scores.

1. 2 Purpose of the study

Weigle (2002) stated, "the score is ultimately what will be used in making decisions and inferences about writers" (p. 108). The purpose of this study is to compare ratings between NES and NNES raters in assessing university students' essays. The present research investigates the differences in scores and the scoring behavior of the two groups of raters. The score differences between the two groups of raters are examined by using the FACETS program. The raters' scoring behavior between the two groups is examined on the basis of the grounded theory. The following is there search questions of this present study :

1. what are the differences between native English speaker (NES) raters and non native English speaker (NNES) raters on assessing written products in both the analytic

scoring method and the holistic scoring method ?

2. what are the assessment categories that NES and NNES raters make most use of in evaluating students' essays as revealed in their written comments? Is there any difference between the two groups?

3. Is there any discrepancy between NES and NNES raters when the criteria of the analytic scoring method are compared with the new set of criteria derived from raters' written comments?

1. 3 Limitations of the study

The present study has a few limitations. One limitation is related to the number of the data used as students' essay writings. In the present study, a total of 39 essays were used to be rated. The sample size was not considered to be sufficient to generalize the result s and to setup a model of rating behavior. Originally, 47 writing samples were supposed to be provided by professors teaching a writing class. However, eight essays were not gathered because of class dropping or being absent on the test day.

The second limitation is that raters' assessment experiences were not controlled. This could act as a variable in this study. However, this was an inevitable choice because they have been assigned to listening, reading and grammar classes from a school and the school have not opened a lot of English writing classes. Accordingly, it seems to be difficult for NNES teachers to have writing assessment experiences in Korean university context.

Another additional limitation is about the way of analyzing and interpreting the qualitative data. In this study, hand analysis was chosen for the qualitative data, which meant there searcher had to read the data, mark them, and divide them into parts by hand (Creswel, 2011). Although it was positive to be able to be close to the data, this was a labor-intensive activity, a time consuming process to commit to, and a lack of visual effect to show the result. Using a specific software program such as *NVivo* (www. qsrinternational. com) would also provide for visual mapping categories identified in the analysis.

1. 4 Chapter Overview

Chapter 2 covers the literature review on which this research is grounded. First partisan overview of performance assessment and writing assessment then, the characteristics of raters in writing assessment are introduced. The section of rater characteristics is divided into five parts: 1) rater reliability and rater bias, 2) NES raters vs. NNES raters, 3) scoring methods, 4) assessment criteria, and 5) Rating experience. Lastly, the two analysis methods used in the present study are reported : many-facet Rasch measurement and grounded theory.

Chapter 3 provides details concerning the data and the method s in this study. It describes participants including raters and students, instruments for gathering the data, procedure for there search, and the way of analyzing the data.

In Chapter 4, the result s of the study are presented in the order of there search questions. Finally, Chapter5reports the discussion and conclusion related to the present English language education context. Additionally, some implications for the study are suggested and the future of research related to rater study in language assessment is presented.

CHAPTER 2 LITERATURE REVIEW

This chapter reviews the literature regarding raters' characteristics and their assessment behavior in scoring students' written compositions. This chapter largely consists of three parts: writing assessment, the characteristics of raters, and analysis method s.

In the first part, this consideration began an overview of performance assessment and writing assessment. The next part covered the characteristics of human raters. Factors that have affected rater characteristics were divided into five aspects concerning this study. The first aspect involved the rater's reliability and bias in writing assessment. Following this, the studies concerning NES and NNES raters were reported, which was very important for this study because the present study aimed to compare their rating characteristics and behavior on writing assessment. Next, two scoring method s were introduced: analytic and holistic scoring methods. Then, the studies of assessment criteria and rating experience were reviewed.

It was a useful review to understand raters' performance. Last part is about methods; many-facet Rasch measurement for quantitative analysis and coding analysis based on grounded theory for qualitative data were introduced.

2. 1 Review of performance -based assessment

The term of performance assessment was synchronously used with alter native assessment (Herman, 1992). The alter native assessment was conducted in applied linguistics (Linn, Baker, &Dunbar, 1991; Wiggins, 1989) However, they have some what different meanings.

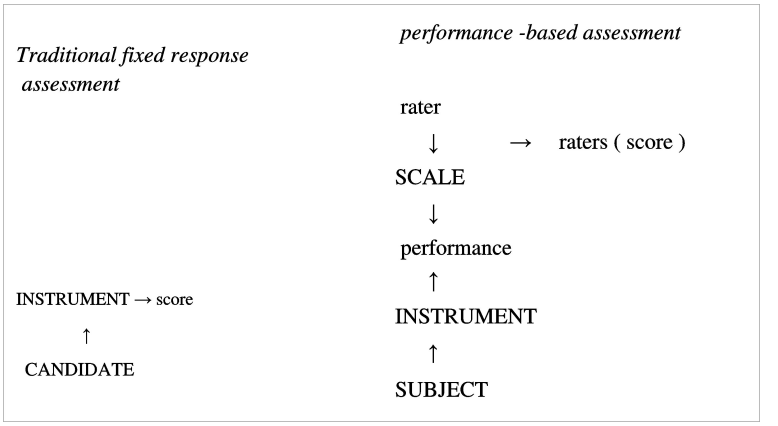
The term alter native assessment is an antithetic concept to the multiple-choice, timed, and one-shot approaches that characterize the standardized assessments. Strictly speaking, traditional assessment should be compared with alter native assessment. The term of

authentic assessment has the idea that assessments should engage students in applying knowledge and skills in the same way they are used outside of school-"real world" (Shohamy, Gordon, & Kraemer, 1992). However, Linn and Gronlund (2000) have labeled the term performance assessment because "it is more descriptive than alter native assessment and less pretentious than authentic assessment " (p. 260). Performance assessment is a broad term which brings both authentic assessment and alter native assessment. Marzano, Pickering, and McTighe (1993) covered a broad scope of the performance assessment in their book as follows:

Performance assessment refers to a variety of tasks and situations in which students are given opportunities to demonstrate their understanding and to thoughtfully apply knowledge, skills, and habits of mind in a variety of contexts. These assessments often occur overtime and result in a tangible product or observable performance. They encourage self-evaluation and revision, require judgment to score, reveal degrees of proficiency based on established criteria, and make public the scoring criteria (p. 13).

As the characteristics of performance assessment suggested by McNamara (1996) have been the most representative, McNamara (1996) compared that assessment with traditional assessment (see Figure1).

Figure 1
characteristics of performance assessment (McNamara, 1996, p. 9)



In comparison with traditional assessment, performance assessment has a very distinct

difference in that presence of a performance and a judging process occur. In traditional fixed response assessment, the score comes from a instrument with a correct answer such as in a multiple choice test. However, the score of performance -based assessment is derived from a combination of rater judgement and interpretation of scale descriptors. In this setting, a new interact ion can be introduced: an interact ion between the rater and the scale, and between the examinee and the instrument. Thus, to assess a performance test fairly, the point of departure should be to identify the rater-scale interact ion and subject-instrument interact ion. Another characteristic of performance assessment has been identified by O'Maly and Valdez Pierce (1996, cited by Brown, 2005). Basically, they were likely to interpret the assessment from the aspect of tasks that students conducted, and suggested 6 characteristics of performance assessment as follows:

1. students make constructed responses.
2. They engage in higher-order thinking, with open-ended tasks.
3. Tasks are meaningful, engaging, and authentic.
4. Tasks call for the integration of language skills.
5. Both process and product are assess ed.
6. Depth of a student' mastery is emphasized over breadth.

The research design of the present study is based on writing performance assessment. Hamp-Lyons (1991) suggested five characteristics of writing performance assessment :1) each examinee writes at least one piece of composition with more than 100words, 2) the examinee is provided with a place within which to create a response to the given text, 3) the composition written by an examinee must be usually read and evaluated by more than two or more raters who have been through some training for the essay evaluation, 4) the evaluation made by raters should have some guide lines such as a set of sample essays, concrete descriptions demonstrating a certain levels or rating scales, and 5) the raters' evaluation to the composition should not be conducted by one rater; other higher or external authority can record and retrieve scores on the test.