



# Native and Iranian teachers' perceptions and evaluation of Iranian students' English essays



Fahimeh Marefat\*, Mojtaba Heydari<sup>1</sup>

Faculty of Persian Literature and Foreign Languages, Allameh Tabataba'i University, South Allameh Street, Saadatabad, Tehran, Iran

## ARTICLE INFO

### Article history:

Received 2 April 2015

Received in revised form 10 October 2015

Accepted 21 October 2015

Available online 9 December 2015

### Keywords:

Writing assessment

Many-facet Rasch measurement

Rater bias

Perception

## ABSTRACT

Past research has shown different reactions by native English speaker (NES) and non-native English speaker (NNS) teachers to student writings. The present study aims to investigate whether NES raters differ from Iranian raters in the severity with which they rate essays. It also sets out to see to what extent such differences reflect the raters' perceptions of what makes a good writing. A total of 144 teachers (NES = 72, Iranian = 72) participated in an online survey, ranking four writing criteria (i.e., content, organization, vocabulary, and grammar) in order of difficulty and importance. Subsequently, 24 of the teachers (NES = 12, Iranian = 12) rated eight essays written by Iranian EFL undergraduates based on IELTS band descriptors and considering the criteria in question. The results of a many-facet Rasch measurement showed that Iranian raters were significantly more severe than NES raters in rating the essays. Also, the results of the survey revealed significant differences between the two groups of raters in their perceptions of the difficulty and importance of organization and grammar. The findings suggest that rater training programs in Iran should aim at deemphasizing the role of grammar while at the same time highlighting the role of inter-sentential discourse.

© 2015 Elsevier Inc. All rights reserved.

## 1. Introduction

In recent years, there has been a growing interest in performance-based assessment which involves rating student writings using a set of predetermined criteria. However, this has raised serious questions regarding the reliability and validity of ratings (Hamp-Lyons, 1990; Henning, 1996; McNamara, 1996; Mehrens, 1992). For example, research has found a great degree of unwanted rater variability (Bachman, Lynch, & Mason, 1995; Eckes, 2005; Engelhard & Myford, 2003; Lumley & McNamara, 1995). Based on O' Sullivan's (2000) study, Shaw and Weir (2007) reported on rater characteristics that might affect rater performance. These characteristics include physical/physiological (e.g., age, sex, etc.), psychological (personality, cognitive style, etc.), and experiential variables. As some of these factors may not easily lend themselves to investigation, previous research has mainly focused on experiential factors including professional background (e.g., Brown, 1995), level of experience (e.g., Cumming, 1990), and linguistic background (e.g., Johnson & Lim, 2009).

Since many teachers and raters of English are non-native speakers (NNS), the role of native English speakers (NES) as the only norm makers in language assessment has been questioned by many scholars (e.g., Kachru, 1997; Taylor, 2006), and much effort (e.g., Lee, 2009; Shi, 2001) has been made to investigate whether NES and NNS raters judge students' performance

\* Corresponding author. Tel.: +98 2188692349; fax: +98 2188690020.

E-mail addresses: [f.marefat110@gmail.com](mailto:f.marefat110@gmail.com) (F. Marefat), [m.heydari.ac@gmail.com](mailto:m.heydari.ac@gmail.com) (M. Heydari).

<sup>1</sup> Tel.: +98 9127425387.

differently. However, most studies on the differences between NES and NNS raters have only dealt with speaking, and only a few have focused on writing assessment (e.g., Johnson & Lim, 2009). Also, the emphasis of such studies has been mainly on the differences in outcomes (i.e., the ratings provided by raters) (e.g., Schaefer, 2008; Shi, 2001) rather than on the differences in beliefs and perceptions. The present study aims to see if there are differences between NES and NNS raters in rating student writings and if such differences could be attributed to the raters' perceptions of the characteristics of a good writing.

## 2. Review of literature

The term bias may refer to every construct-irrelevant variability which affects a rater's decision about a test taker's performance. Bias studies look for relationships between score variability and other variables such as raters' expertise, language background, and gender. According to Schaefer (2008), an important part of bias analysis is to search "for unexpected interactions between rater judgments and test takers' performance or other facets" (p. 467). Numerous studies have been conducted on rater bias in assessing both speaking and writing.

Regarding oral assessment, Reed and Cohen (2001) surveyed the relevant literature on the characteristics of raters which might result in score variation. The characteristics reported by them included: native/non-native, raters' occupation, gender of raters, and personality interaction between rater and candidate. They suggested that, (a) further research should encompass both qualitative and quantitative methodologies to investigate rater behavior, and (b) new types of rating scales should be constructed to reflect test constructs more explicitly. Later, Lumley (2002) used a think-aloud protocol to investigate the decision making process of four trained and experienced raters and found that raters sometimes apply the contents of the scale in quite different ways. He concluded that even after rater training, raters can decide to pay attention to some scales more than the others, and justify their impressions of the text in terms of the parameters signified in the scale and rater training sessions. In another study, O'Loughlin (2002) investigated the impact of gender of candidates and raters on an IELTS oral interview using discourse analysis of the interviews and a multi-faceted Rasch analysis of test scores. The results from both analyses revealed no significant impact of gender on the oral interview.

More related to the current study, another line of research has focused on rater characteristics in assessing students' essays. In a review of 20 major empirical studies on ESL writing assessment, Huang (2009) reported several variables that might affect the rating of students' essay. The variables included rater background, mother tongue, previous experience, amount of prior training, and types and difficulty of writing tasks. In addition, Haswell and Haswell (1996) emphasized the role of gender as another variable leading to rater bias. Based on their empirical study, they reported that raters underrated the essays written by students of their own sex.

Standard practice in bias studies is to have NES and NNS raters rate some essays in order to compare the ratings of the two groups to find potential bias pattern(s). Some researchers have employed an analytical scoring method (e.g., Kobayashi, 1992; Schaefer, 2008), whereas others have used the holistic method of scoring (e.g., Brown, 1991). Still, some have used both methods concurrently in an attempt to compare the resultant ratings (e.g., Connor-Linton, 1995; Lindsey & Crusan, 2011; Song & Caruso, 1996).

In analytical scoring, writings are rated according to their surface-level (e.g., grammar and vocabulary) and discourse-level features (e.g., content and organization). Regarding surface-level features, research has mostly shown that NES and NNS teachers do not give similar scores, with NES teachers being generally more tolerant of student errors (Brown, 1991; Hyland & Anan, 2006; Song & Caruso, 1996). Hyland and Anan (2006), for example, gave a correction task to two groups of raters, Japanese and English EFL teachers, and found that NNS teachers were more severe in rating grammatical errors. A similar severity was reported by Connor-Linton (1995), who compared the American and Japanese instructors' ratings and found that NNS teachers tended to focus on surface-level features, whereas NES teachers were more sensitive to inter-sentential discourse features. In addition, more recently, Lee (2009) studied the rating behavior and perceptions of five Korean English teachers and five NES teachers who graded 420 essays written by Korean college students. Using the data from test scores and a survey questionnaire, Lee found that Korean raters' were inferior in measuring linguistic components, and were more severe in scoring grammar, and sentence structure.

As for the discourse-level features, research suggests that NES raters tend to emphasize the importance of inter-sentential discourse, organization, content, and overall criteria more than do NNS raters (Hughes & Lascaratou, 1982; Santos, 1988). For example, Lee (2009) found that NES raters are more severe with content and overall criteria than are NNS raters. Similar results have been reported for native speaker raters of other languages when assessing writings in their respective languages. For instance, Kuiken and Vedder (2014) reported that Dutch and Italian raters attach more importance to discourse features than to surface features. It should also be noted at this point that some studies (e.g., Lee, 2009, on Korean raters) have shown NNS raters to be generally more severe with organization. Furthermore, Kobayashi (1992) found Japanese NNS teachers to be less lenient in rating clarity and organization than were NES teachers.

Employing the holistic method of scoring to compare NES and NNS raters has yielded contradictory findings. On the one hand, Song and Caruso (1996) showed NES raters to be more lenient than ESL teachers in their holistic scores. However, Machi (1988) compared the essay-grading behavior of NES and Japanese-speaking teachers of English and found insignificant differences between the two groups. Brown (1991), similarly, reported that there are insignificant differences between NES and ESL teachers in their ratings of compositions written by college students. Likewise, Shi (2001) found no significant difference between the holistic ratings of NES and Chinese raters. However, she posited that the two groups of raters may have arrived at similar scores from different viewpoints. Employing both analytic and holistic methods, Lindsey and Crusan

Download English Version:

<https://daneshyari.com/en/article/344200>

Download Persian Version:

<https://daneshyari.com/article/344200>

[Daneshyari.com](https://daneshyari.com)