Assessing Writing xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Assessing Writing

A Many-Facet Rasch analysis comparing essay rater behavior on an academic English reading/writing test used for two purposes

Sarah Goodwin

Georgia State University, United States

ARTICLE INFO

Article history:
Received 23 January 2016
Received in revised form 15 July 2016
Accepted 20 July 2016
Available online xxx

Keywords:
Second language writing assessment
Many-Facet Rasch measurement
L2 writing raters
Factors affecting writing scores
Rater variability

ABSTRACT

Second language (L2) writing researchers have noted that various rater and scoring variables may affect ratings assigned by human raters (Cumming, 1990; Vaughan, 1991; Weigle, 1994, 1998, 2002; Cumming, Kantor, & Powers, 2001; Lumley, 2002; Barkaoui, 2010). Contrast effects (Daly & Dickson-Markman, 1982; Hales & Tokar, 1975; Hughes, Keeling, & Tuck, 1983), or how previous scores impact later ratings, may also color raters' judgments of writing quality. However, little is known about how raters use the same rubric for different examinee groups. The present paper concerns an integrated reading and writing test of academic English used at a U.S. university for both admissions and placement purposes. Raters are trained to interpret the analytic scoring rubric similarly no matter which test type is scored. Using Many-Facet Rasch measurement (Linacre, 1989/1994), I analyzed scores over seven semesters, examining rater behavior on two test types (admissions or placement). Results indicated that, of 25 raters, five raters showed six instances of statistically significant bias on admissions or placement tests. The findings suggest that raters may be attributing scores to a wider range of writing ability levels on admissions than on placement tests. Implications for assessment, rater perceptions, and small-scale academic testing programs are discussed.

1. Introduction

This paper presents an investigation into rater behavior on an integrated reading and writing task used for both university admissions and placement purposes. Using Many-Facet Rasch measurement (Linacre, 1989/1994), I describe how raters interpreted scores for two different academic English testing populations. This analysis is intended to contribute to the body of research on second language writing in assessment contexts and also to the ongoing research and validation for the in-house English as a second language (ESL) testing program from which the data came. The testing program coordinator must train human raters of our reading and writing test to ensure that the scores they assign are satisfactory measures of writing performance. It is important to remember that examinee scores on performance assessments are mediated through human raters. In other words, ratings themselves are not a direct representation of the quality of examinee writing, because the rater's experiences and judgments play a role (Wind & Engelhard, 2013). Because of the considerations mentioned here, an investigation is necessary that examines raters' scores of writing quality and what effects there may be on the assignment of scores.

E-mail address: saregoodwin@gmail.com

http://dx.doi.org/10.1016/j.asw.2016.07.004
1075-2935/

2. Background to the study

Scores assigned by writing test raters may be impacted by various characteristics of raters' background, such as their experiences with grading writing exams, teaching writing, prior language learning, and so forth (Barkaoui, 2010; Cumming, 1990; Cumming, Kantor, & Powers, 2001; Lumley, 2002; Vaughan, 1991; Weigle, 1994, 1998, 2002). Weigle (1998) and Lumley (2002), in investigations of raters of second language writing, both note the complex nature of the rating process. The textual features of an essay, the wording of the rating scale, and all of the impressions readers bring with them – as well as the potential interaction of these elements – can have an effect on raters' perceptions of writing, and thus on the scores they assign.

Raters' judgments of what they have read may be impacted by the quality of the samples they have previously rated, reflecting possible contrast effects on scores. On average, raters scored an average-quality composition preceded by high-quality samples lower than when it followed lower-quality exemplars (Daly & Dickson-Markman, 1982; Hales & Tokar, 1975; Hughes, Keeling, & Tuck, 1983). Additionally, Spear (1997) found that two preceding samples of contrasting quality created stronger biasing effects than just one sample on scores assigned to later pieces of writing. Although raters may be instructed in their rater training to, for example, not compare samples to one another during the scoring process, prior investigations illustrate that quality distinctions among essays may have an impact on later scores assigned.

Murphy and Yancey (2008) even go so far as to say that "[rater] variability represents an underlying disagreement about the nature of the construct underlying the assessment" (p. 369), although they are referring to writing assessment in general rather than specifically to second language writing tests. McNamara (1996), however, states that it may be suitable "to accept variability in stable rater characteristics as a fact of life, which must be compensated for in some way, either through multiple marking and averaging of scores, or using the more sophisticated techniques of multi-faceted analysis" (p. 127). Hence, variability does not necessarily indicate that raters are arriving at vastly different judgments of writing quality. Even if raters have a keen understanding of the writing assessment construct, it may not be essential to have raters be trained to be completely rigid in their interpretations of texts, as some variability is bound to occur.

Rasch (1960/1980) measurement theory has been employed as a method for examining rating quality in writing assessments (e.g., Engelhard, 2002, 2013; Linacre, 1989/1994; Wind & Engelhard, 2013; Wolfe, 2004). The Many-Facet Rasch measurement (MFRM) approach to monitoring rater performance considers facets that may have an impact on scores. Facets can be components such as raters, rating scales, or examinees, and these are plotted onto a common interval scale along with scores that raters assign (Bachman, 2004; Eckes, 2009; Lim, 2011). More detail about the interval scale is presented in the methodology section of this paper. In order for meaningful information to be drawn from MFRM measures, two assumptions need to be made: the data should fit the model, and the test should measure a single, unidimensional construct (Bond & Fox, 2007; Eckes, 2008; McNamara, 1996).

Other investigations into second language writing assessment have employed MFRM methods to investigate rater traits. Weigle (1998) used MFRM to examine differences in rater severity and consistency before and after training, finding that rater training helped boost scorers' intra-rater reliability (internal consistency). Also examining rater harshness/leniency, as well as accuracy/inaccuracy and centrality/extremism, Wolfe (2004) employed MFRM to "control for error contributed by systematic variability between both items and raters" (p. 42). While Weigle (1998) used a pre- and post-test design and Wolfe's (2004) study was a one-time snapshot of raters, Lim (2011) examined rater consistency and severity longitudinally, over 12–21 months, for novice and experienced raters. Investigating various writing rater effects may contribute to both improved scoring and rater training. Moreover, it can provide information for the validation of writing assessment rating scales (Harsch & Martin, 2012; Knoch, 2011; Shaw & Weir, 2007).

The body of existing research underscores that raters must contend with a number of variables while rating, participating in tasks that require judgments drawing on various sources of information. The monitoring of rater quality with regard to how raters use rating scales is thus of importance to contribute to the validity and reliability of tests. The current study concerns raters, examinees, and a scale for the essay component of an academic English reading and writing test administered to two test-taker populations.

3. Context for the study

The academic English proficiency test from which the data are drawn is an examination given to ESL examinees for college admissions decisions and ESL course placement recommendations. Designed by a team of ESL instructors, content-area instructors, and language assessment professionals, it is intended to reflect the types of tasks students will need to perform in university contexts. It consists of an integrated reading and writing task, a multiple-choice listening comprehension section, and a multiple-choice reading comprehension section. New graduate students also sit a face-to-face oral interview.

The integrated reading and writing section consists of a short-answer and an essay component, requiring examinees to read two source texts and synthesize them in their writing responses (the short-answer component) as well as reading and responding to an argumentative writing prompt (the essay component). The short-answer component is scored in content and language, and the essay component is rated for content, organization, accuracy of language (grammar and vocabulary), and range and complexity of language. The essay portion, which is a timed argumentative writing sample, will be the focus of the present investigation (its rubric can be found in Appendix A). Each analytic rubric category (content, organization, accuracy of grammar/vocabulary, and range/complexity of grammar/vocabulary) is scored along a 10-point scale using

Please cite this article in press as: Goodwin, S. A Many-Facet Rasch analysis comparing essay rater behavior on an academic English reading/writing test used for two purposes. *Assessing Writing* (2016), http://dx.doi.org/10.1016/j.asw.2016.07.004