Assessing Writing 25 (2015) 38–54

Contents lists available at ScienceDirect

Assessing Writing

ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study

Paula Winke [a,*], Hyojung Lim [b]

[a] Michigan State University, United States
[b] Hankuk University of Foreign Studies, Seoul, South Korea

ARTICLE INFO

Article history:
Received 5 January 2015
Received in revised form 11 May 2015
Accepted 12 May 2015
Available online 4 June 2015

Keywords:
Cognitive processing
Rater effects
Analytic scoring
Inter-rater reliability
Essay rating
Rubric use

ABSTRACT

We investigated how nine trained raters used a popular five-component analytic rubric by Jacobs et al. (1981; reproduced in Weigle, 2002). We recorded the raters' eye movements while they rated 40 English essays because cognition drives eye movement (Reichle, Warren, & McConnell, 2009): By inspecting to what raters attend (on a rubric), we gain insights into their thoughts. We estimated inter-rater-reliability for each subcomponent. Attention (measured as total eye-fixation duration and eye-visit count, with the number of words per subcomponent controlled) was associated with inter-rater reliability: *Organization* (the second category) received the most attention (slightly more than the first, *content*). *Organization* also had the highest inter-rater reliability (ICC coefficient = .92). Raters attended least to and agreed least on *mechanics* (the last category; ICC coefficient = .85). Raters who agreed the most had common attentional foci across the subcomponents. Disagreements were directly viewable through eye-movement-data heatmaps. We discuss the rubric in terms of primacy: raters paid the most attention to *organization* and *content* because they were on the left (and read first). We hypothesize what would happen if test developers were to remove the least-reliable (and right-most) subcomponent (*mechanics*). We discuss rubric design as an important factor in test-construct articulation.

© 2015 Elsevier Inc. All rights reserved.

1. ESL essay raters' cognitive processes: an eye-movement study

In this study, we examine how raters cognitively process an analytic rubric while rating English-as-a-second-language (ESL) essays to understand more fully construct-irrelevant variation in essay-test scores. Only a small number of empirical studies have examined the cognitive processes of raters during essay-rating. As we will review below, the researchers (Cumming, Kantor, & Powers, 2002; Lumley, 2005; Sakyi, 2000) have mainly used think-aloud protocols alongside questionnaires and/or interviews to investigate the raters' cognitive processes. In this study we built upon this prior research and recorded essay raters' eye movements while they used an analytic rubric. We did this to derive a finer picture of how raters attend to the various components of a rubric. We coupled the quantitative eye-movement data with qualitative reviews of heatmaps and gaze plots to better understand whether different patterns in processing the rubric are related to unreliability in test scores.

* Corresponding author at: Department of Linguistics and Languages, Michigan State University, B252 Wells Hall, 619 Red Cedar Road, East Lansing, MI 48824, United States. Tel.: +1 517 353 9792.
E-mail addresses: winke@msu.edu (P. Winke), hyojunglim.sls@gmail.com (H. Lim).

http://dx.doi.org/10.1016/j.asw.2015.05.002
1075-2935/© 2015 Elsevier Inc. All rights reserved.

1.1. Essay rating processes

Researchers have outlined what they believe is a common procedural path that raters follow when they rate second-language (L2) essays. Cumming et al. (2002) suggested the following prototypical sequence for raters who rate Educational Testing Service's TOEFL (http://www.ets.org/toefl) essays, which are scored holistically. First, raters scan and quickly assess an essay. Then, they reread the script for further comprehension and for judgment. Finally, raters justify their scores and reinterpret their judgments. Similar patterns were described by Lumley (2002), who investigated how experienced raters used an analytic scale to score the writing section of the Australian immigration English test. The three-stage model of the rating processes, first suggested by Freedman and Calfee (1983), appears to be applicable to both holistic and analytic scoring, regardless of the purpose of test administration. At the individual level, however, variations in rating behaviors have been constantly observed through think-aloud or questionnaire data (Cumming et al., 2002; Eckes, 2008; Wolfe, 1997; Wolfe & Kao, 1996). Thus, it is not certain how applicable the three-stage model is in operational testing: Individual raters may or may not conform to this general model.

Variations in rating processes are often ascribed to individual differences in raters' attentional focus. Raters may focus on different features in test takers' essays when scoring, or they may weight the different scoring categories differently (Cumming et al., 2002; Eckes, 2008; Orr, 2002). Raters might consider external features that are not even described in a rubric, such as the length of the essay or even the test takers' handwriting (Barkaoui, 2010a; Lumley, 2005; Vaughan, 1991). Reasons for such differences in raters' cognitive processes seem to vary greatly. The variation may depend on whether the essays are scored holistically or on an analytic scale. Below we look at these two conditions in turn.

Several researchers (Cumming et al., 2002; Shi, 2001; Cumming, 1990) have investigated raters' processes in rating essays holistically. When scoring holistically, raters are basically in control of their own thought processes because they only need to decide upon one general score, and not several individual ones (as when they use an analytic rubric). By interviewing raters, Cumming (1990) found that more experienced raters employed a larger and more varied number of essay-rating criteria during holistic rating, whereas novice raters scored based on only a few component skills with which they were more familiar (based on their prior teaching and editing experience). In Cumming et al. (2002), non-native essay-raters reported (on questionnaires) that they paid more attention to *language*, whereas native-speaking essay raters noted they equally valued *language*, *rhetoric*, and *ideas*. In a similar vein, Shi (2001) discovered that nonnative raters focused more on *content* and *organization*, whereas native essay-raters emphasized *language use* when holistically scoring. These researchers found that raters' professional backgrounds, language backgrounds, and scoring experience impact the way the raters read essays and assigned holistic scores. But such findings may not be unexpected given that in holistic rating, the aspects on which the essay raters assign scores is largely un-prescribed (see Knoch, 2009).

Different from holistic rating, when using an analytic rubric, raters are provided with a prescribed list of linguistic features, and they must rate them individually. An analytic rubric can be seen as a map guiding the thought-processes involved in scoring (Knoch, 2009). Barkaoui (2010a) and Lumley (2005) investigated analytic essay-scoring processes by interviewing raters while they rated. Barkaoui (2010a) explained that the effect of rating-scale type was greater than that of rater experience on raters' decision-making behaviors. Holistic scales tended to encourage more use of interpretation strategies (e.g., strategies to comprehend an essay), whereas analytic scales elicited more judgment strategies (e.g., strategies to formulate a rating) and self-monitoring foci (e.g., monitoring for personal bias). With holistic scoring, raters read or focused on the essay for a longer period of time, which Barkaoui explained was needed for the raters to rationalize their score assignments. With analytic scoring, on the other hand, raters' attention was more directed to the rating scale for the articulation and justification of their scores. Consistent with Lumley (2002), with an analytic rubric, Barkaoui explained, raters attended to all scoring categories outlined in the rubric, whereas with holistic rubrics, raters focused on what they thought was important. Though further evidence is needed, such observations may offer reasonable accounts for why analytic scales contribute to increased intra-rater reliability[1] and why holistic scales are relatively more subject to a halo effect.[2] For instance, novice raters in Barkaoui (2010a) focused heavily on a number of specific linguistics features with holistic scoring, whereas with analytic scoring, they were better able to evenly distribute their attention to all scoring categories.

Individual differences among raters seem to be persistent, despite the provision of well-developed rating scales and rater training. According to the rater-type hypothesis (Eckes, 2008), even experienced raters using analytic scales will not agree on the aspect of writing they focus on most. Eckes conducted a survey-based study in the context of a test of German as a foreign language and investigated which criteria raters considered most important. Results supported the rater-type hypothesis; even with appropriate training and experience, raters held different views concerning criterion importance. Eckes classified the raters into six different types; a subsequent correlation analysis implicated the effect of raters' age, living-abroad experience, and language-learning backgrounds. Older raters thought of the syntax criteria (e.g., vocabulary,

[1] Intra-rater reliability is a method of estimating the consistency of judgments by calculating a correlation coefficient of two sets of scores produced by the same rater for the same group of students (Brown, 2005, p. 288).

[2] A halo effect occurs when raters fail to discriminate between a number of conceptually distinct categories, but rather rate a candidate's performance on the basis of an overall impression, so that raters award the same score across a number of different rating scales (Knoch, Read, & von Randow, 2007, p. 27).

Download English Version:

https://daneshyari.com/en/article/344208

Download Persian Version:

https://daneshyari.com/article/344208

Daneshyari.com