Title of Project:
The Effects of Primacy on Rater Cognition:
An Eye-tracking Study

Researcher:
Laura Ballard
Michigan State University
lauradballard@gmail.com

Research Supervisor:
Dr. Paula Winke
Michigan State University
winke@msu.edu

Project Summary

Motivation for the Research
Rater scoring has an impact on performance test reliability and validity. Thus, there has been a continued call for researchers to investigate issues related to rating (Crusan, 2015). Myford (2012) exhorts researchers and test designers to “do all that [they] can to help ensure that the ratings that raters assign are accurate, reliable, and fair” (p. 49). Second language testing researchers are committed to this goal and have been researching the various facets that affect test scoring processes for years (Cumming, Kantor, & Powers, 2002; Eckes, 2008; Kondo-Brown, 2002; Lumley, 2002; Orr, 2002). In second language writing assessment, such emphasis on investigating the scoring process and how raters arrive at particular scores have been seen as critical “because the score is ultimately what will be used in making decisions and inferences about writers” (Weigle, 2002, p. 108).

In the current study, I answer the call for continued research on the rating process by investigating rater cognition in the context of writing assessment. Research on raters’ cognitive processes “is concerned with the attributes of the raters that assign scores to student performances, and their mental processes in doing so” (Bejar, 2012, p. 2). A theme central to rater cognition is the way in which raters interact with rubrics. Only by understanding this interaction will test designers be able to improve rubrics, rater training, and test reliability and validity (Barkaoui, 2010). Performance test validity is tied to raters and rubrics, in particular, because there are certain propositions that must be counted as true in order for scores to be considered valid (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, see Standard 6.9). For example:

• “Raters attend to the criteria included in the rubrics when making their judgments (i.e., they are using appropriate criteria when they are assigning their ratings).
• Raters use the categories on the rubrics in the intended manner, applying the rubrics consistently and accurately to judge each performance (or product)” (Myford, 2012, pp. 48-49).

In this study, I focused on rater-rubric interactions, which continue to be of interest because, despite rater-training efforts, variance in rater behavior and scores persist (Lumley & McNamara, 1995; McNamara, 1996; Weigle, 2002; Weir, 2005) which may lead to reliability problems. Though the goal of rater training is to give raters a common understanding of the rubric criteria and to help raters converge
on a common understanding of scoring bands (Bejar, 2012; Roch, Woehr, Mishra, & Kiesczynska, 2012), many studies on rater behavior have shown that raters do not always use rubrics in a consistent way (i.e., they have low intra-rater reliability). Raters do not consistently score (i.e., they have low inter-rater reliability), and they do not use the same processes to arrive at a given score (Cumming, Kantor, & Powers, 2002; Eckes, 2008; Kondo-Brown, 2002; Lumley, 2002; Orr, 2002). As Winke and Lim (2015) suggested, one potential explanation for rater behavior and problems with inter-rater reliability may be the primacy effect. The primacy effect is a psychological phenomenon that shows that the positionality of information in a list (e.g., a rubric) affects a listener’s or reader’s assignment of importance to that information (Forgas, 2011). This seems particularly relevant for helping to explain how raters pay attention to rubric criteria. Primacy may have a potential impact on inter-rater reliability on analytic rubrics. No researcher, however, has directly investigated the role of primacy in rater cognition and its potential effects on rater scoring; however, Winke and Lim (2015), posited that primacy effects were observable in their study. Thus, I investigated primacy effects in relation to rater-rubric interactions, and I examined whether they affect behavior, such as mental-rubric formation, attention to criteria, and rater scoring, when raters use an analytic rubric.

Research Questions

1. To what extent do raters show evidence of ordering effects in their mental-rubric formation after rater training?
2. To what extent do raters show evidence of ordering effects through their rubric usage during rating?
3. To what extent are raters’ scores impacted by ordering effects?

Research Methodology

I employed a mixed-methods within-subjects design and included eye-tracking methodology, criteria importance surveys, criteria recall tasks, decision-making process outlines, and rater interviews. Thirty-one novice raters were randomly assigned to two groups, who, for counterbalancing purposes, were trained on two rubrics in two phases. The rubrics were a standard rubric (SR; from Polio, 2013) and a reordered rubric (RR; identical to the SR, except with categories appearing in a mirrored order to the SR). In Round 1, raters trained on one of the two rubrics and rated the same 20 essays using the rubric. The second round took place five weeks after the completion of the first. In Round 2, raters trained on the alternate rubric and re-rated the same 20 essays. Throughout the two rounds, I utilized several data-collection tools to investigate raters’ cognition and behavior related to their rubric of training. Using Criteria Importance Surveys (CIS), I examined raters’ beliefs about category importance. From the Criteria Recall Tasks (CRT), I examined raters’ recall of the descriptors in each rubric category. With eye tracking methodology, I recorded the raters’ focus on the rubric criteria during essay rating to uncover how raters used the rubric criteria based on the position of the categories. Finally, from raters’ essay scores, I examined the raters’ scoring consistency and severity for each rubric category.

Summary of Findings

The multiple data measures tell the same story: as novice raters train on a new rubric and assign scores using the individual categories on the rubric, the raters’ behavior pertaining to the outermost positions (e.g., left-most and right-most) seems most susceptible to ordering effects. That is, the findings of this study show that the category position affected the raters’ beliefs about what criteria are the most and least important when scoring an essay, how many descriptors raters were able to recall from a category, how much attention raters paid to a category on the rubric while rating, and how
severely raters scored a given category. Additionally, the findings provided evidence that there was an interplay between the category types and category positions, resulting in either more pronounced primacy effects or leveling effects for individual rubric categories. Perhaps most importantly, there was evidence of a halo effect, in which the first category affected raters’ scoring severity in the subsequent categories.

Implications

Based on the findings of this study, it would be beneficial for test designers to carefully consider the layout and ordering of analytic rubrics used in operational testing. Rubric designers could leverage ordering effects to their benefit by fronting any categories that are typically seen as less important or have lower interrater reliability scores. Test designers may also want to consider making word counts similar across categories (as done by Polio [2013] in her paper on revising the Jacobs et al. [1981] rubric) and striving for clarity and precision in each individual descriptor in order to reduce the amount of rater interpretation needed for a descriptor, as requested by Knoch (2009).

Given that raters may become more and more entrenched in their beliefs and scoring patterns when rating over long periods of time, test designers could also consider creating an online rater-training and scoring platform (see Knoch, Read, & von Randow, 2007; Wolfe, Matthews, & Vickers, 2010) which would encourage raters to pay equal attention to each rubric category. One example may be a digital platform that presents raters with a randomized, forced order of training, norming, and scoring. For each essay, the platform could randomly prompt raters to score a given category, only allowing raters to score one category at a time and input scores for the category appears on the screen. This may reduce rater’s conditioning to attend most to certain categories while least to others. Additionally, many researchers advise having two raters score each essay (Elder, Barkhuizen, Knoch, & von Randow, 2007; Lumley & McNamara, 1995; Marzano, 2002; McNamara, 1996), and if raters trained and scored on categories in a random order, then pairs of raters would provide a more balanced scoring scheme and would be an additional step to mitigate any effects of primacy on scoring.

In the case that rating programs intend certain categories to be more important, those categories should be left-most, and training should indicate that the left-most categories are more important and explain why. I suspect that this is being done subconsciously in rating programs that use analytic rubrics. The rater training most likely has the new raters learn about the categories in the order they are presented (from left to right on the rubric). The rater trainers most likely work through sample scoring scenarios using the rubric from left to right, and may even unintentionally spend more time explaining the left-most categories. This ordering may have an effect on mental rubric representation, how raters view the importance of the categories, and how well certain categories are used over time. This study shows that ordering effects are real. Rater training programs now need to use that information to better design rating programs such that any ordering effects are intentional and to the betterment of the program, or the category ordering needs to be controlled so that ordering effects will not take hold and be detrimental to the rating program over time.
References


