Same text different processing? Exploring how raters' cognitive and meta-cognitive strategies influence rating accuracy in essay scoring

Jie Zhang[*]

Shanghai University of Finance and Economics, 503 Red Tile Building, 777 Guoding Rd., Yangpu District, Shanghai, 200433, China

A R T I C L E   I N F O

A B S T R A C T

Rater variability has long been a concern in performance assessment. Among the various factors leading to variability, raters' thinking processes may play an important role in shaping the rating outcomes. However, few studies have attempted to link raters' cognition with rating accuracy. Thus, this study aims to explore how raters' use of cognitive and meta-cognitive strategies might influence rating accuracy in the context of College English Test Band 4 essay scoring in China. Thirteen CET4 raters were classified into ACCURATE and LESS ACCURATE groups based on their rating accuracy, which was operationalised as the match between the ratings obtained from the target rater and those obtained from an expert panel measured by the Many-Facet Rasch Model analysis. Concurrent think-aloud was utilised to elicit the cognitive and meta-cognitive strategies the raters employed during the rating process. The thinking processes of the two rater groups were compared both quantitatively and qualitatively. The results suggest that more accurate raters are better at integrating information from target essays and are more self-conscious about their rating accuracy. Findings are expected to explain to some extent why some raters are effective in decision-making and generate more accurate ratings, which will have important implications on rater training.

© 2015 Elsevier Inc. All rights reserved.

1. Introduction

Performance assessment (typically writing and speaking assessment) is a direct measure of learners' communicative language competence. Its authenticity and positive washback on teaching and learning has been widely acknowledged. Despite its advantages, the elicitation of complex constructed response inevitably requires human raters to evaluate the language use performance of examinees. Rater variability has long been regarded as one of the most significant sources of measurement error and a potential threat to the reliability and fairness of performance assessment. Therefore, raters' behavior has been gaining tremendous attention in language testing research. (Bachman, Lynch, & Mason, 1995; Baker, 2012; Bonk & Ockey, 2003; Brown, Iwashita, & McNamara, 2005; Eckes, 2008, 2012; Kuiken & Vedder, 2014; Myford & Wolfe, 2000; Weigle, 1998; Wolfe, 1997).

Rating constructed response is in nature a cognitively complex decision-making process, which involves information acquisition and processing. Myford and Wolfe (2003: 387) described the possible processes going through raters' minds

* Tel.: +86 13764193852.
E-mail addresses: zhang.jie@mail.shufe.edu.cn, zhangjie617@gmail.com

as follows: 'The persons who render their judgments function as information processors. They engage in a highly sophisticated, complex mental process to arrive at their decisions-observing, recalling information from memory storage, and then organising, combing, weighing, and integrating that information to draw inferences about individuals.' Many studies in language testing literature explored how raters process the information acquired from the target performance and the mental procedures they follow when making scoring decisions. (see review below)

However, previous studies mostly focused on the thinking processes without linking raters' cognition with rating outcomes. Hence, it remains unclear whether the differences in raters' thinking processes have impact on the rating quality and more importantly whether some processes may yield better and more accurate scoring decisions. To bridge the gap, this study aims to explore how raters' use of cognitive and meta-cognitive strategies might influence their rating accuracy in the context of College English Test Band 4 (CET4) essay scoring in China. The participants included 13 CET4 raters and two quality control experts. The raters were classified into ACCURATE and LESS ACCURATE groups based on their rating accuracy, which was operationalised as the match between the ratings obtained from the target rater and those obtained from an expert panel measured by the statistical indices from the Many-Facet Rasch Model (MFRM) analysis. Concurrent think-aloud was utilised to elicit the cognitive and meta-cognitive strategies the raters employed in making scoring decisions. The thinking processes of the two rater groups were compared both quantitatively and qualitatively. Findings are expected to explain to some extent why some raters are effective in decision-making and as a result generate more accurate ratings, and this will have important implications on rater training.

2. Rater cognition studies in writing assessment

Since the early 1990s, the research on rater variability has been witnessing a cognitive shift. More studies began to focus on the decision-making processes involved in scoring (Barkaoui, 2007, 2010; Connor-Linton, 1995; Deremer, 1998; Erdosy, 2004; Hamp-Lyons, 1991; Lumley, 2002; Vaughan, 1991; Wiseman, 2012). Typically employing a qualitative and process-oriented approach, in particular verbal protocol analysis, researchers identified a number of decision-making strategies or reading styles utilised by raters when evaluating essays. Vaughan (1991) asked nine experienced raters to think aloud while scoring six essays using a six-point holistic scale. She listed the essay elements which were commented on more than three times by each rater and identified five reading styles according to their use of certain rating strategies. Specifically, the 'single-focus approach' mainly focuses on a particular aspect such as the pass/fail decision, the 'first impression dominates approach' relies predominantly on the raters first and often intuitive impression of the target essay and the 'two-category strategy' takes into account mainly two categories of textual features, while the 'grammar-oriented reader' focuses primarily on grammar and the 'laughing rater' is more likely to give affective responses. Vaughan concluded that, 'Despite their similar training, different raters focus on different essay elements and perhaps have individual approaches to reading essays' (p.120). Deremer (1998) conceptualised the rating process from a problem-solving perspective and showed that the ill-structured nature of rating activities defied a unified solution procedure. She explored the task elaboration of three highly experienced raters (i.e., the way these raters defined and implemented a rating task) and identified three types of task elaboration: simple recognition task elaboration, rubric searching elaboration, and complex recognition elaboration. These types represent three different foci, namely, general impression scoring, text-based evaluation, and rubric-based evaluation. Milanovic, Saville, and Shen (1996), by employing multiple data collection methods, like group interviews, retrospective written reports and introspective verbal reports, also identified four marking approaches used by raters, which were referred to as principled two-scan/read, pragmatic two-scan/read, read through, and provisional mark. They also noted that raters weighed in radically different manners the 11 essay elements they mentioned in their verbal reports. More recently, studies went further to categorise raters into different types based on certain criteria. For instance, Eckes (2008, 2012) categorised raters according to the weight they attached to different rating criteria, namely, the Syntax Type, the Correctness Type, the Structure Type, the Fluency Type, the Non-fluency Type and the Non-argumentation Type, while Baker (2012) classified raters on their decision-making style profiles in five dimensions, namely, rational, intuitive, dependent, avoidant and spontaneous.

The above studies simply described raters' thinking processes in general. Lumley's study (2002) focused on how raters utilised rating scale in decision making. He collected the concurrent think-aloud protocols of four trained and experienced raters to determine how they understood and applied the features of an analytic rating scale and investigated the role of the rating scale in their rating process. In developing coding schemes for data analyses, Lumley focused on the raters' interpretations of scoring categories, the difficulties they encountered in the rating process and the strategies they employed to handle such challenges. Additionally, he found that the raters exerted conscious efforts to relate to the rating scale; however, there was always a tension between the rules and the raters' intuitive impressions due to the scale's inability to cover all eventualities in written texts and the occasional conflicts between specific textual features and the scale descriptors on a certain score level. Furthermore, the inadequacy of the rating scale forced raters to develop various strategies to cope with problems in the rating process, which would inevitably lead to indeterminacy and confusion. Lumley thus concluded that individual self-monitoring processes come into play when raters try to reconcile their overall impression of a text, the specific, hard-to-evaluate textual features and the rating scale descriptors.

However, the aforementioned studies generally described and categorised raters' mental processes in vague and broad terms. This may well conceal the subtle individual differences among raters. Moreover, these studies did not differentiate raters' cognitive strategies from meta-cognitive strategies. As defined in general psychology, cognitive strategies generally