Title of Project:
Investigating the Combined Effects of Rater Expertise, Working Memory Capacity, and Cognitive Functionality on the Scoring of Second Language Speaking Performance

Researcher:
Qie Han
Teachers College, Columbia University
qh2139@tc.columbia.edu

Research Supervisor:
Dr. James Purpura
Teachers College, Columbia University

TIRF Research Topic Investigated:
Language Assessment

---

Final Report

Motivation for the Research
In second language (L2) speaking assessment, raters can significantly affect test validity due to rater variability, conceptualized as a source of construct-irrelevant variance in scores caused by individual differences in raters’ characteristics rather than examinees’ ability. To improve the validity of interpretations and decisions based on scores, we must investigate what rater characteristics are likely to contribute to rater variability. One of the most frequently examined rater characteristics is rater expertise. Rater expertise refers to raters’ relevant knowledge and experience in L2 performance assessment, and it has been found to associate with different levels of scoring performance. Despite its effect on scoring performance, rater expertise is likely not the only rater characteristic that contributes to rater variability. The possible effects of two other rater characteristics (i.e., raters’ working memory capacity [WMC] and cognitive functionality), also need to be investigated. The reason is that, on the one hand, WMC has been found to impact, either independently or in combination with expertise, a variety of complex cognitive performances, such as language comprehension and acquisition. On the other hand, the cognitive functionality of raters, defined in terms of the strategies related to information processing, underlies the different stages of raters’ scoring process and the mental processes associated with those stages. Through an exploration of the combined effects of these rater characteristics on scoring performance, we can strengthen our assumptions about what rater-associated factors lead to rater variability, thereby shedding light on rater selection, training, and scoring practices. With improved scoring performance, test scores can capture a more accurate portrayal of L2 learners’ speaking ability.

Research Questions
Recently, an increasing amount of research has been conducted to investigate the effect of rater expertise on scoring performance in L2 performance assessment. However, few studies have
investigated the joint influence of rater expertise and raters’ cognitive characteristics on scoring performance. The current study was thus designed to address this gap by exploring the following three main research questions:

1. What were the joint influences of rater expertise and WMC on raters’ scoring performance?
   1-1. What were the relative contributions of rater expertise and WMC to scoring performance?
   1-2. What possible interaction was there between rater expertise and WMC in their joint influence on scoring performance?
2. What strategies did raters use during their scoring process?
3. How did the expert and novice raters differ in the patterns of strategy use?

Research Methodology
To address these questions, the current study employed a mixed-methods research design to examine the combined effects of three rater characteristics, i.e., rater expertise, WMC, and cognitive functionality, on the scoring of L2 speaking performance. Ninety raters were recruited on a voluntary basis from major universities in the United States and one large testing organization in the United Kingdom to participate in a multi-stage research project, where the raters filled out a rater background survey designed to measure their rater expertise, scored 27 test takers’ speech samples from the Aptis speaking test, re-scored 10 of those samples at least a week later, and completed a cognitive task (i.e., a listening span) that measured their verbal WMC. The quantitative data collected from these sessions included the 90 raters’ holistic ratings, which were analyzed using many-facet Rasch measurement analysis to calculate three scoring performance indices that respectively reflected the raters’ scoring accuracy, severity, and consistency. In addition, the raters’ responses to the rater background survey were used to calculate a composite rater-expertise score, and the raters’ scores for the listening span were used as measures for their verbal WMC. Then, a hierarchical regression analysis was performed to explore the joint influences of rater expertise and WMC on the three aforementioned aspects of scoring performance. Subsequently, six raters (three experts and three novices) were randomly selected from the 90 to participate in a cognitive lab session, where the raters verbally reported their thoughts and mental processes while scoring three speech samples. The raters’ verbal reports were transcribed and analyzed in terms of the strategies that they had used while scoring and how the expert and novice raters differed in strategy use. The results of the quantitative and qualitative data analyses were integrated to reveal the combined effects of the three rater characteristics (i.e., rater expertise, WMC, and cognitive functionality) on scoring performance.

Summary of Findings
The results from the quantitative data analyses (mainly hierarchical regression analysis) have demonstrated that rater expertise significantly predicted scoring accuracy. This result seems to align with previous research indicating that more proficient raters are more accurate and appropriate in their application of the rating scale than less proficient raters (Davis, 2012; H. J. Kim, 2011), especially if the scale is holistic (Barkaoui, 2010). This finding also seems to corroborate previous research which has found significant contributions of expertise to complex cognitive performance in various domains (e.g., Alptekin & Erçetin, 2011; Hambrick & Engle, 2002; Joh & Plakans, 2017; Payne et al., 2009, to name a few), and may be explained by theories
about expertise and expert cognition, such as experts’ deeper, better-organized and more functional knowledge representation (Chase & Simon, 1973a, 1973b; Chiesi et al., 1979; de Groot, 1978; Simon & Chase, 1973; Spilich et al., 1979) and superior skills of using pre-established knowledge structures during task performance (Endsley, 2018; Ericsson & Kintsch, 1995; Feltovich et al., 1984). Despite the significant effect of rater expertise on scoring accuracy, the effect size of rater expertise found was a bit small, probably due to the fact that rater expertise was measured in the current study as raters’ general, assessment-related experiences rather than test-specific knowledge and skills.

What was a bit surprising in my findings was that WMC was not found to significantly predict scoring performance. This result may be explained by the fact that the current study was designed to reduce raters’ cognitive load during the scoring process, which might have mitigated the contribution of WMC to scoring performance. Specifically, the current study attempted to make raters’ scoring process as authentic as possible to real-life scoring practices. As a result, not only were the raters allowed to replay spoken responses during the scoring process, but were also given constant access to the scoring rubric. Replays during the scoring process could have reduced the essentialness of a larger WMC. In the same way, the raters’ WMC could have contributed more to scoring performance if the raters were not given constant access to the scoring rubric, which could have forced them to rely more on WM to process and integrate the rubric criteria with their pre-established conceptualization of L2 speaking ability. As a result of these factors, the effect of WMC on scoring performance may have been moderated. As for the interaction between rater expertise and WMC, the non-significant interaction effects found in the current study seem to support the independent influences model, which hypothesizes no interaction between expertise and WMC in their joint influences on cognitive performance (Alptekin & Erçetin, 2011; Hambrick & Oswald, 2005; Payne et al., 2009).

The results from the qualitative analyses identified fourteen major (meta)cognitive strategies that the raters had reported using during the scoring process. Amongst these strategies, three were most frequently reported by the raters: applying criteria from the rubric, evaluating, and retrieving relevant response information from WM. These three strategies were the building blocks of a rater’s scoring process and were thus most commonly used by the raters. Moreover, the raters were frequently found to have used these strategies as a strategy cluster, which seems to align with the information processing model and the different stages of scoring proposed by Purpura (2012) and Bejar (2012).

Based on the strategies identified from the raters’ verbal reports, differences in the quantity and quality of strategy use between the expert and novice raters were explored. The results have revealed that the two groups of raters used similar ranges of strategies. However, the expert raters used eight strategies considerably more frequently than the novice raters, which include applying criteria from the rubric, comprehending, evaluating, linking to prior knowledge or experience, noticing, reasoning, reflecting (especially on one’s own scoring performance), and retrieving relevant response information from WM. These differences, as discussed in detail in my dissertation, can mostly be attributed to the experts’ higher level of expertise in L2 speaking assessment. Not only were differences found in the quantity of strategy use between the expert and novice raters, but the two groups of raters were also found to differ in the quality of strategy use. The results have shown that the expert raters had demonstrated more success and competence in using strategies and strategy clusters than the novice raters. These differences may be explained in light of the characteristics of expertise, such as experts’ better recognition and recall of new materials involving domain-relevant information and better integration of new
information within a coherent and interconnected framework of existing knowledge and information to make useful inferences (Chiesi et al., 1979; Spilich et al., 1979). Relevant details can be found in the results and discussion sections of my dissertation.

**Implications**
The present study has a number of implications for the assessment of L2 speaking performance. First, the significant effect of rater expertise on scoring accuracy seems to support our assumption about the contribution of raters’ accumulated experience in L2 assessment to scoring performance. However, the low effect size of rater expertise has raised questions about measuring rater expertise as raters’ general, assessment-related experiences rather than test-specific knowledge and skills. This finding may warrant a reconsideration of our commonly-used criteria (e.g., general rating and teaching experiences) for rater selection and categorization in existing L2 research and rating practices, especially when it comes to predicting raters’ scoring performance for a specific test. Second, although WMC is a basic ability required for performing a variety of complex cognitive tasks, the relative contribution of WMC to scoring performance, compared to that of rater expertise, was found much less appreciable in the current study. This finding seems to corroborate existing findings about the more predominant role of expertise in cognitive performance (e.g., Hambrick & Engle, 2002; Joh & Plakans, 2017; Payne et al., 2009). From another perspective, it also seems to support the effectiveness of our commonly-used measures (e.g., giving raters opportunities to re-listen to responses and constant access to the scoring rubric) to reduce raters’ cognitive load in real-life rating practices. Lastly, differences in the expert and novice raters’ strategy use, both in terms of the frequency of using certain strategies and the efficiency of using strategies in general, were discovered. These differences can not only help explain scoring performance on a deeper, cognitive-processing level (Purpura, 2014) but also indicate the potential of training novice raters on strategy use similar to an expert rater’s during the scoring process to improve scoring performance.
References


Purpura, J. E. (2012). *What is the role of strategic competence in a processing account of L2 learning or use?*. Paper presented at the American Association for Applied Linguistics Conference, Boston, MA.


