Title of Project:
The Impact of Rater Experience and Essay Quality on Rater Behavior and Scoring

Researcher:
Ozgur Sahan
Çanakkale Onsekiz Mart University
ozgursahan66@hotmail.com

Research Supervisor:
Prof. Salim Razi
Çanakkale Onsekiz Mart University

---

Final Report

Motivation for the Research
Students’ English language writing performance is assessed for a variety of purposes at Turkish universities. First, students must pass high-stakes entrance and exit exams with a writing section as part of the one-year, intensive English language preparatory program (ELPP), which students are required to complete before starting their English-medium departmental studies. Students’ writing performances are also evaluated throughout their university education in order to track their progress in different courses, such as academic writing or advanced writing. Further, some universities require a good command of EFL writing as a prerequisite for exchange programs like Erasmus+ because students pursue their studies with such exchange programs in English. The task of preparing exams for the aforementioned purposes typically falls to ELPP testing units or course lecturers. However, scoring procedures do not always follow formal, predetermined steps, such as training and calibrating raters to rate the essays reliably. As such, different assessment protocols are implemented at different institutions and within the same institution. Therefore, there is a need for a standardized and sound assessment system in order to provide students with fair scorings.

Although several factors can contribute to score variations in writing performance assessments, rater-related factors can be considered central to the subjectivity of writing assessment. As one of the rater features, previous rating experience is attributed to ensuring fair judgment, placing expert scorers in a superior position throughout the evaluation processes. Yet, expertise in assessing writing does not necessarily guarantee reliable scores. Therefore, it is essential to understand the differences and commonalities in raters’ reactions to essays of different qualities in order to understand better the variability of ratings. To this end, this research study focuses on two factors—scorers’ rating experience and essay quality—to investigate their impact on the variability of EFL essay scores and rating behaviors in Turkish tertiary-level education. Examining assessment problems commonly seen at institutional and national levels, this research gains
significance by investigating two main sources of variation in EFL writing assessment to establish meaningful and generalizable measurements that should be relevant beyond individual contexts.

**Research Questions**

The main purpose of this study was to investigate the impact of rater experience and essay quality on rater behavior and essay scores. Adopting a mixed-methods research design, the variability of ratings assigned to the essays were examined from quantitative and qualitative perspectives with the following two sets of questions.

From the quantitative aspect, the first set of questions were as follows:

1. Are there any significant differences among the analytic scores of the low- and high-quality EFL essays?
2. Are there any significant differences among the analytic scores assigned by raters with varying previous rating experience?
3. What are the sources of score variation that contribute most (relatively) to the score variability of the analytic scores of EFL essays?
4. Does the reliability (e.g., dependability coefficients for criterion-referenced score interpretations and generalizability coefficients for norm-referenced scores interpretations) of the analytic scores of raters differ based on their amount of experience?

Moreover, based on the qualitative data, the following questions were asked:

5. How do raters make decisions while rating different quality EFL essays analytically?
6. How is rating experience related to EFL raters’ decision-making processes and the aspects of writing they attend to?

**Research Methodology**

Employing a mixed-methods research approach, this study used a convergent parallel design (Cresswell, 2011). Within this research design, the qualitative and quantitative strands of research are concurrent during the process of data collection and within the overall interpretation of results but independent of each other during the data analysis phase. Thirty-three raters voluntarily participated in this study. The participants were based in 16 different state universities in 15 different cities. All the participant raters were professionals in the field of interdisciplinary English language teaching, learning and assessment, and regular employees at the School of Foreign Languages (SFL), Foreign Languages (FL) Department, or English Language Teaching (ELT) Department at state universities in Turkey. These 33 raters were all graduates from different ELT and ELL departments in Turkey, and they have the same L1 background (Turkish). The participants varied in their professional experience in teaching and assessing EFL writing. Based on their reported rating experience, participants were divided into three groups: low-experienced (n = 13), medium-experienced (n = 10), and high-experienced raters (n = 10).

The data collection tools used in this study consisted of 50 EFL essays, a 10-point analytic scoring rubric, verbal protocols, written score explanations, and a background questionnaire. Each of these materials was designed carefully before the commencement of the main data collection phase. The essays were used to obtain both qualitative and quantitative data. Before data collection, the essays were evaluated by three independent expert raters for quality division, resulting in two sets of high-quality and low-quality texts. Raters were asked to score the essays using an analytic rubric to which they had been oriented and to provide three written score explanations for each...
essay regarding their judgments. Added to that, each rater employed think-aloud protocols (TAPs) while scoring 16 essays pre-determined by the researcher. Raters were trained on how to conduct TAPs through detailed guidelines and a sample TAP video prior to data collection. Before submitting their essay packs to the researcher, raters filled out a background questionnaire as well. A total of 9,900 scores (1,650 total scores and 8,250 sub-scores), 446 think-aloud protocols, and 5,425 written score explanations were obtained from the participants. The analysis of quantitative data relied on generalizability (G-) theory approach as well as descriptive and inferential statistics; qualitative data were analyzed through deductive and inductive coding.

**Summary of Findings**

Raters showed statistical differences in their scores assigned to high-quality and low-quality essays. In other words, all raters were able to distinguish low-proficient student authors from their high-proficient peers. Raters varied from one another in their ratings based on their previous rating experience. High-experienced raters and low-experienced raters displayed statistically significant differences in their total ratings of low-quality essays. Furthermore, statistically significant differences were observed between their sub-scores assigned to the mechanics component of the low-quality essays. When the scoring pattern across experience groups was examined, a positive relationship between the average scores and the amount of rater experience was observed in that more experienced raters gave higher scores to the essays than low-experienced raters did.

In addition, G-theory analysis revealed that the variance due to raters was considerably high when the ratings of high-quality and low-quality essays were evaluated separately. However, the score variability due to raters was much smaller collectively, indicating that raters showed greater differences in terms of leniency and severity within each essay quality than in the overall mixed-quality set. Fourth, an almost perfect degree of inter-rater reliability was achieved within each rater group for low-quality and mixed-quality (high- and low-quality papers together) essays, and D-studies showed that a lower number of raters would still produce scores with an acceptable level of dependability index. However, the reverse is true for high-quality essays in that low-dependability coefficients were found across the three rater groups, and only if the number of raters were increased unreasonably would reliable scores be obtained for high-quality essays.

When it comes to scoring behaviors, raters displayed different decision-making strategies based on essay quality and rating experience. More experienced raters were more positive compared to less experienced raters, leading to higher essay scores respectively. Generally, raters used more interpretation strategies than judgement strategies. Raters focused more on style, grammar, and mechanics when rating low-quality essays but more on ideas, rhetoric, and their general impression of the essay when rating high-quality essays. Added to that, medium- and high-experienced raters displayed similar decision-making behaviors, while low-experienced raters differed slightly from these two more experienced groups. Low-experienced raters used more interpretation strategies than their more experienced peers whereas medium- and high-experienced raters employed judgment strategies more frequently than raters with less experience did. Medium-experienced and high-experienced raters tended to employ the same strategies while rating essays of both low- and high-quality. For both low- and high-quality papers, the low-experienced raters seemed to rely on more language-focused strategies, particularly with respect to mechanics. Across experience groups, raters displayed more language-focused strategies—such as considering punctuation, spelling, and syntax—for low-quality essays than high-quality essays.
Implications
The findings of this study underline the need for detailed and continuous rater training even for raters with extensive rating experience. In this way, scoring gaps can be reduced between raters. Traditional rater training models can be revisited as the findings suggested that score variations between raters may be related to differentiation in certain sub-scores of writing (e.g. mechanics component), as certain raters (e.g. low-experienced raters) prioritized strategies related to such components (e.g. consider spelling and punctuation) in their think-aloud protocols and the written explanations. As such, developing a rater-training model that shifts raters’ focus to all aspects covered by the scoring criteria instead of emphasizing certain traits such as grammar, content, or organization might help ensure intra- and inter-rater reliability. That is, a strategy-based rater-training model built upon the most commonly used decision-making strategies may lead raters to think similarly while evaluating EFL compositions, thus resulting in more consistent scores.

Another implication addresses double-grading protocols for institutional assessment. While many institutions use double-grading, protocols for matching the rater pairs are rarely considered. Given that high-experienced raters were found to be more lenient compared to their less-experienced peers, language programs can consider matching relatively high-experienced and relatively low-experienced raters as double-grading pairs. In other words, if two high-experienced raters are paired, they may be more likely to give higher scores to a certain writing performance, while the same essay might receive a considerably lower score if the grading is conducted by two less-experienced raters. Matching relatively high- and low-experienced raters together could compensate for these effects in double-grading situations.

Although analytic rubrics are considered more reliable and advantageous than holistic scoring, the findings showed that score variations could be observed during analytic evaluation. As such, rather than using traditional holistic and/or analytic scoring scales, developing a clear and user-friendly scale with more detailed descriptors might be helpful to reduce inconsistencies between raters. Added to that, context-bound scoring scales can be developed with specific consideration of the local, cultural, and institutional dynamics.
References


BTU SFL (2014). *Holistic scoring scale*. Bursa Technical University School of Foreign Languages, Turkey.


