Teaching English Language Journal
ISSN: 2538-5488 – E-ISSN: 2538-547X – http://tel.journal.org

Please cite this paper as follows:
Ahmadi, A., Tavazoei, M. (2025). Rater training through eye-tracking: A case-study of a novice rater. *Teaching English Language, 19*(2), 511-539. https://doi.org/10.22132/tel.2025.472698.1668

Research Paper

Rater Training Through Eye-Tracking: A Case-Study of a Novice Rater

Mina Tavazoei
Ph.D. Candidate, Department of English Language and Literature, Faculty of Foreign Languages, University of Shiraz
Alireza Ahmadi [1]
Professor, Department of English Language and Literature, Faculty of Foreign Languages, University of Shiraz

Abstract
This study centered around the notion of rater training with the help of eye-tracking systems. A novice rater participated in a rater training program which was informed by tracking the rater's eye movements. Immediately after rating a sample of essay in each session, the rater was provided with eye-tracking feedback in the form of a heat-map produced based on his eye movements. The heat map was discussed to help the rater understand his behavior during the rating and to pinpoint which rubric descriptors and essay parts the rater noticed more while rating. The findings revealed that in the early sessions, the rater was influenced by the primacy effect; that is, he was mostly focusing on the two first criteria (content and organization). Furthermore, initially, he had struggles deciding on a band score and dedicated considerable attention to the scores rather than descriptors. However, after sessions of training the rater appeared to modify his behavior and tried to focus on all criteria and the equivalent descriptors. The findings can assist rater trainers in organizing rating programs more effectively by employing eye-tracking systems to scrutinize raters' behavior.

***Keywords**:* Rater Training, Essay Scoring, Eye-Tracking, Cognitive Process, Novice Rater

[1] Corresponding author: arahmadi@shirazu.ac.ir

Received: August 10, 2024
Accepted: August 22, 2025

1. Introduction

Performance-based assessment is critically significant in the realm of language evaluation, primarily due to its concentrated examination of human behavior within genuine contexts and the assessment process carried out by a trained rater (Weigle, 2002). Growing employment of performance-based assessment has shifted scholars' focus to raters and the rating mechanism (Schaefer, 2008). Due to their heavy reliance on human rating (Jin & Eckes, 2022), the quality of ratings is of paramount importance in these assessments (Hamp-Lyons, 2007; Johnson & Lim, 2009). Owing to this dependability on raters' judgments, they are also called rater-mediated assessments (Eckes, 2015; Engelhard Jr, 2013). It is noteworthy that rating extends beyond performance-based assessments; it is utilized in diverse contexts and for various purposes, including educational, clinical, business, and industrial settings (Myford & Wolfe, 2003). Moreover, despite the advent of automated scoring, it is naïve to assume that human scoring, or "expert judgment," has become obsolete (Bejar et al., 2006). The continued relevance of human evaluators in the rating process is underscored by the complexity of this decision-making task, which demands significant cognitive effort (Cumming et al., 2002; Luoma, 2004). This highlights the critical role that human judgment plays in ensuring accurate and nuanced evaluations.

Due to its cognitively-demanding nature, there are several investigations on raters' cognition to gain deeper insight into the rating process (e.g., Bejar et al., 2006; Cumming et al., 2002; Knoch, 2009; Lumley, 2002, 2005; Vaughan, 1991). Suto (2012) argues that cognitively speaking, rating expertise lay its foundation on three discrete competencies 1)

comprehensive realization 2) strategic selection and 3) successful application. Rater cognition plays a crucial role in rater training, as it directly influences the decision-making process (Vaughan, 1991). To enhance the validity of test scores, it is essential for raters to maintain consistency in their cognitive approaches. However, this consistency can paradoxically result in rater variability, which may stem from differing conceptualizations of the same factors among individual raters (Wolfe, 1997). Therefore, it is important to recognize that rating should not be oversimplified as merely assigning scores to students (Weigle, 2002).

Despite the use of identical judgment criteria, rater variability remains a challenging issue in human scoring (Bejar et al., 2006). It has always been questioned whether the final scores are true scores or the result of interaction between different parties (Weigle, 1999). To address the external and internal factors that can lead to inaccurate ratings and to minimize rater subjectivity, rater training is an essential step.

Rater training can be defined as the process of familiarizing raters with rating criteria and providing them with the opportunity to score essays to develop a shared understanding of the rating procedure (Elder et al., 2007). Therefore, it involves three phases of familiarization, radicalization (practicing sample essay scoring), and discussion (Lane & Stone, 2006). According to Winke and Brunfaut (2021), the result of rater training would be raters' consistency across different tests and rubrics. The training will lead to better rater behaviors and expertise and reduce rater effects in the end (e.g., Saito, 2008; Winke & Brunfaut, 2021; Youn, 2018). Furthermore, rater training is a tool to modify raters' presuppositions about learners' characteristics and task features to lessen variabilities among raters. However, the objective should not be to rely on machine rating to decrease discrepancies among raters, but rather to ensure they have a solid

understanding of the rubric. While some researchers argue that extensive rater training can be resource-intensive and may not completely eliminate subjective biases (Lumley & McNamara, 1995), rater training still remains as a crucial component (Winke & Brunfaut, 2021).

Generally, the efficiency of rater training sessions is not guaranteed. Some studies suggested rater training could be beneficial (e.g., Erlam et al., 2013; Shohamy et al., 1992); while in some studies its effectiveness remained inconclusive (e.g., Wang & Engelhard Jr, 2019; Wind, 2019a). Moreover, if at first, rater training proved to be efficient the results were not long-lasting (Knoch, 2011). Due to discrepancies in study results, think-aloud verbal protocols (TAPs) were used in the rater training sessions to gain a better understanding of the cognitive processes that take place while rating. There are several studies that employed TAPs in combination with other instruments to gain deeper comprehension over raters' cognitive processes (e.g., Lumley, 2005; Barkaoui, 2010). However, using TAPs in rater training sessions has proven to be insufficient on its own, as their effectiveness can be influenced by the rater's experience and the rubrics used. TAPs placed additional demands on the raters, potentially disrupting the rating process, impacting their evaluations, and providing limited insights into the rating process (Lumley, 2005). Barkoui (2011) stated that TAPs are "indicators of cognitive processes, rather than direct evidence of their full realizations" (p. 27). Hence, following the process of development in organized sessions for each rater is crucial for better comprehension of the efficiency of the designed rater training sessions and raters' progress along the cognitive development path (Li et al., 2019). Considering essay scoring which is cognitively complex, and the use of a multi-trait rubric for scoring makes it harder and more cognitively demanding for raters (Luoma, 2004), a more sophisticated tool should be employed to better conduct the study. Another

useful instrument that can be employed for rater training to compensate for the limitations of previously employed instrument (i.e., TAP) is eye-tracking. Since the data provided by eye-tracking is rich and comprehensive and result in natural moment-to-moment data sources (Conklin & Pellicer-Sánchez, 2016), this is called a methodology delves deeper into cognitive efforts rather than just "scratching the surface of the cognitive process" (Low & Aryadost, 2021, p.14).

This study seeks to enhance the understanding of the essay scoring process through the use of eye-tracking systems and to evaluate the effectiveness of these systems in rater training programs. It was conducted by scrutinizing one rater's eye movements during rater training sessions while scoring English essays using an analytic rubric. It is believed that in comparison with holistic rubrics, analytic rubrics are designed to allocate equal portion of attention to each criterion separately (Winke & Lim, 2015). Given that prior research has highlighted the limitations of Think-Aloud Protocols (TAPs), this study employed eye-tracking systems as a methodological tool to reveal the underlying cognitive processes that remain obscured in TAPs. The rater's heat-map for each session is recorded for analysis of how eye movements changed during rater training sessions.

2. Literature Review

According to Myford and Wolfe (2003), judges act as information processor tools. This information processing role of raters can be considered a process that involves problem-solving and being engaged in interaction with several parties, rather than simplistically assuming raters as an entity to read a text and give a score (Deremer, 1988). Raters gather information from their environment, organize them, then integrate them to conclusions, and occasionally they record their judgments via rating scales (Deremer, 1998). This judging process is most often threatened by several sources of error. If

these sources of error are treated as insignificant, they will deliberately affect rating validity (Bachman, 1990; Myford & Wolfe, 2003). Additionally, these processes and the reason behind measurement errors should be clarified and evaluated to result in validity and reliability refinement (Lumely, 2005). Raters, themselves, might affect the process of rating based on their leniency or severity, central tendency, and halo effect (Myford & Wolfe, 2003, 2004). So, the issue of rater variation in essay rating is undeniable (Linacre, 2002). Ample researches have illuminated construct irrelevant factors that might affect scoring (Barkaoui, 2007; Bejar et al., 2006; Cumming et al., 2002; Eckes, 2008; Engelhard Jr, 1994; Hoyt, 2000; Johnson & Lim, 2009; Lumley, 2005; McNamara, 1996; Myford & Wolfe, 2003, 2004; Shi, 2001; Weigle, 2007).

Raters may be influenced by various construct-irrelevant factors, such as prior experience (Cumming et al., 2002), rubric criteria and students' level of proficiency (Schaefer, 2008), task types (Shin, 2009), or even the type of rubric (Barkaoui, 2010). More surprisingly, even the co-construction of a rubric did not exhibit a shared and consistent understanding of the criteria among raters (Deygers et al., 2015). Even, the employment of e-rubrics induced more inconsistency in rating for a rater (Erguvan & Aksu-Dunya, 2021). In a nutshell, rater variability inevitably exists, and rater training sessions should be conducted as a remedial course to compensate for the detrimental effects of rater variability (Harsch & Martin, 2012; Myford & Wolfe, 2003) on essay scoring.

Diederich et al. (1961) pioneered the importance of rater training to improve behaviors and reduce rater effects. In the same vein, Jacobs et al. (1981) recommended training with a rating scale to ensure raters understand criteria and avoid bias. However, the effectiveness of rater training has been debated, with some studies showing positive results (e.g., Cumming, 1990;

Elder et al., 2005; Knoch et al., 2007; Shohamy et al., 1992; Weigle, 1994, 1998) and others finding no reduction in bias (e.g., Elder et al., 2007; Knoch, 2011; Lumley & McNamara, 1995). For instance, after training, raters differed in degree of severity, and they had different interpretations of rating criteria (Lumley, 2005). Or, some raters depicted new areas of bias after rater training sessions (Elder et al., 2007). Moreover, rater training did not eliminate raters' bias across male and female test takers (Wind, 2019b). Due to the complexity of the process of rating, more complicated arrangements and instruments are required for rater training to be constructive (Cumming, 1990).

Various techniques such as TAPs, Rasch measurement, G theory, correlation with external variables, regression, and ANOVA across different constructs like speaking, reading, and writing are used to assess ratings (Wind &Peterson, 2018). It is noteworthy that, the effectiveness of previously used instruments (i.e., TAP) is being questioned. Neither Rasch measurement nor ANOVA alone uncovered the hidden cognitive processes during training. Furthermore, TAPs proved inefficient due to their added demands on raters, which could disrupt the rating process, affect evaluations, and offer limited insights (Lumley, 2005). Eye tracking is a valuable tool for rater training which can addresses the limitations of TAPs. It involves examining cognitive processes underlying eye movements on a computer screen (Godfroid, 2019). Eye tracking provides detailed and extensive data, allowing for a deeper exploration of cognitive processes (Conklin & Pellicer-Sánchez, 2016). Language assessment is also influenced by the introduction of eye-tracking systems.

Eye trackers have proven to be effective and beneficial when concentrating exclusively on language evaluation and essay scoring. For instance, Winke and Lim (2015) in an eye-tracking study illustrated how

raters were influenced by the criteria mentioned first in the rubric. They recorded the eye movements of nine raters and their essay scoring process based on Jacobs et al.'s (1981) rubric. Findings indicated that raters paid more attention to components that were introduced early (organization and content) and their fixation duration decreased moving toward the right side of the rubric (language use and mechanics). Interestingly, the heat-map exhibited the reason for agreement between teachers who agreed the most in their scoring. To compensate for the issue of primacy effect, Ballard (2017) employed a revised version of Jacobs et al. (1981) rubric in two formats (original and mirrored version). Several instruments were used in this study including: eye tracking, stimulated recalls, and interviews to ensure a rigid and precise process of data analysis. Overall, it was concluded that items' positions in rubrics influence raters' processes of rating. Eye trackers have also been used to scrutinize the effect of a rater's background (composition teacher or ESL writing teacher) in writing assessment (Eckstein et al., 2018). The findings suggest that an individual's disciplinary background influences the scoring process, thereby leading to varied comprehension of the text.

The studies reviewed above emphasize the importance of rater training and highlight the subjectivity that raters often face during evaluations. Given the limitations of TAPs—which can provide limited insights, are intrusive, and cognitively demanding (Lumely,2005)—eye-trackers present a novel solution. Their exclusive features, such as objective data collection, real-time analysis, and non-intrusiveness, can effectively enhance rater training by fostering a shared understanding of the rating process.

Previous research has examined raters' behaviors after training by tracking eye movements during a single rating session (e.g., Winke & Lim, 2015) or by focusing solely on the rubric to better understand how raters allocate their attention while scoring (e.g., Ballard, 2017). However, there is

a paucity of research on whether and how the results of eye tracking could be employed to train raters. As such, this study aimed at taking the initiative to probe into the potential of eye tracking for rater training.

3. Methodology

3.1 Design

This study, benefited from a qualitative design. Case study research is conducted upon a rater and the results of eye-tracking data, mainly heat-maps, are used to evaluate the rater's behavior and eye movements during rater training sessions. According to Ary et al. (2019) case studies are valuable tools to provide both emic and etic perspectives. Furthermore, being pluralistic, descriptive and heuristic are major characteristics of case studies (Ary et al., 2019).

3.2 Participants

A 21-year-old undergraduate male student of English language and literature who had the experience of teaching English for one year voluntarily took part in this study. He could be considered a novice rater as he had no rater training before and his rating experience was limited to the rating of his students in classroom samples in one term. Furthermore, he had no experience of rating samples with the rubric used in this study. He was ensured that his data would only be employed for research purposes, he would remain anonymous, and he could withdraw from the study at any time.

3.3 Materials and instruments

3.3.1 Student essays

For this study 12 essays with seven different prompts written by female and male students who were between B1 and B2 level were used. The topics were based on content covered in Top Notch series books (Saslow & Ascher,

2015) which were taught in a private language institute in Shiraz. The essays were selected from various prompts to compensate for the effect of prompt on rating process. The prompts were describing a hero, pros and cons of living in an urban area, a successful business and its feature, two different holidays, who a procrastinator is and the following consequence, a mystery and its theory and the ways to overcome a weakness. It is noteworthy that, to better follow the participant's progress and to control the role of topic on the scoring process, just the first and the last essays used in eye-tracking sessions had similar prompts. Additionally, two of the essays were used to train the rater in group training sessions, and two other essays were employed in pre- and post-tests. To better record each participant's eye movements, the essays were not more than one page (between 146 -231 words). They were all in typed version in Times New Roman font style and in size of 12; with a double-space distance from the prompt which was located at the top of the page. Furthermore, they were among low- mid- and high-scoring essays which were scored by the experienced raters of the institute in advance.

3.3.2 Rubric

The rating scale that was employed in this study was Jacobs et al. (1981) rubric. This rubric contains five sub-categories of content, organization, vocabulary, language use, and mechanics, with different weightings. Previous studies using eye tracking for essay rating have also used this rubric (e.g., Winke & Lim, 2015), or in some cases they modified it for their studies (e.g., in Ballard's 2017 study to compare analytic and holistic rubrics). Results depicted that this rubric proved to be useful and produced necessary eye-tracking metrics as expected. Since the production of eye-tracking metrics is highly contingent upon the designed tasks; if the metrics are not met with proper results, in the end the task design must be blamed. This rubric is

believed to be one of the "most widely used analytic scales" which is frequently used in the ESL context (Weigle, 2002, p.115). Furthermore, according to Janssen et al. (2015), this rubric benefits from strong construct validity regarding predefined goals.

3.3.3 Eye tracking

Eye trackers are tools that are employed for better comprehension of cognitive processes involved in different areas of SLA (Godfroind et al., 2020). The objective-based assessment of eye movement provides researchers with rich natural data for further analyses (Conklin & Pellicer-Sánchez, 2016; King et al., 2019). In eye-tracking studies, eye movement is recorded and stored through the employment of specific hardware and software and analyzed based on researchers' predefined areas of interest (AOIs) (Godfroid, 2019). More importantly, the layout of the stimulus and predefined AOIs should be the same during the study (Conklin & Pellicer-Sánchez, 2016). Due to the not-interfering nature of eye-tracking systems, they can be employed to capture readers' attention while reading a text (Conklin & Pellicer-Sánchez, 2016; Godfroid & Spino, 2015) or in other words "looking into the minds of subjects" (Rayner, 1978, p.618). Generally, these systems rely on two underlying presuppositions 1) more cognitively demanding tasks require more fixation time and 2) an item that is fixated on is the item that is considered and noticed (Stewart et al., 2004). In addition to quantitative data, eye trackers provide qualitative data, like heat maps and scan paths. Heat-maps that represent areas with large and a small number of fixations, and scan path that shows the order of fixations and saccades. The range of sampling rate for these systems is between 50 HZ and 2000 HZ (King et al., 2019) which depends on the purpose of the study. For this study, a head-mounted FOVE0 Specs with a sampling rate of 120 HZ and automatic 2-point calibration, and 1.15° degrees of error was used. As this study aimed

at focusing on sentences, and not specific words, as Areas of Interest (AOIs) this sampling rate was appropriate. Eye movements were recorded binocularly, and the participant sat approximately 600 mm from a computer monitor with a resolution of 1280×1960. For the analysis, Gazealytics as a web-based visual eye-tracking analytics toolkit that was developed by Chen et al. (2023) was used.

3.4 Data collection

The participant was requested to join a training session in order to acquire knowledge on essay scoring and rubric content. The training was mostly focused on rating, the steps to be followed in rating, different rating errors, the structure of the selected rubric (Jacobs et al., 1981) and the explanation of unknown concepts to the participant. Following this, the rater was asked to rate three sample essays using Jacobs et al. (1981) rubric. The rater's scores were compared with those of an experienced rater and his justifications for scoring were discussed.

Following the training session, he was required to score another sample essay. Since he demonstrated no problems with the rubric, he entered the eye-tracking phase of the study. This stage included eight sessions of essay scoring. It is worth noting that during this phase, the setting was carefully observed for the eyes to be followed accurately by the eye-tracking system. First of all, dim light was provided in a quiet room, the height of chairs and camera setup was controlled, and stimulus (rubric) maintained the same character including, font, size, layout, and background color for each essay scoring session. To check the usability of the task, a pilot study was conducted first to see whether heat-maps would be produced or not. In case of any problems with this phase, the task was changed and redesigned. Furthermore, the stimulus was repositioned in terms of left or right direction during the study to prevent participants' directional bias. To begin the study,

the first researcher accompanied the participant individually in the laboratory with an eye-tracking system and its connected monitor. Then, the participant took part in calibration process that was defined by following two green dots on the screen. This was how the validity of the system through a calibration phase before each essay scoring was checked. This procedure was repeated to fix all fixation points accurately and precisely, since any evidence of eye drifts should be corrected at the beginning of the study (Conklin & Pellicer-Sánchez, 2016). After the calibration phase, the participant started reading the text and scoring the essay, while simultaneously his eye movements were recorded and observed by the first researcher. Finally, the rater's heat map concerning each rating attempt was produced by Gazealytics. The heat-maps that were immediately produced after each scoring attempt, made the foundation of the subsequent discussion regarding the rater's essay scoring process. That is, immediately after rating a sample of essay in each session the rater was provided with the eye-tracking feedback in the form of a heat-map produced based on his eye movements. The heat map was discussed to help the rater understand his behavior during rating and to pinpoint which rubric descriptors and essay parts the rater noticed more or considerably ignored while rating. When the eye-tracking phase was conducted for all sessions, the participant was required to take part in a post-test session to score another essay without any eye-tracking systems. Overall, the rater took part in 11 sessions of training, including one session of group training, two sessions of individual training without the employment of eye tracker and eight sessions of individual training with eye-tracking systems. Figure one is a schematic representation of the steps carried out in this study.

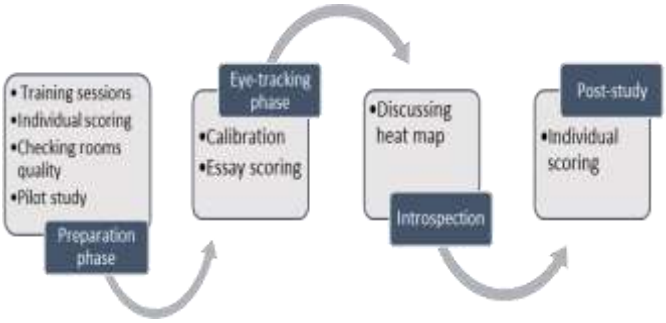

Figure 1. Data Collection Steps

4. Data analysis

In order to analyze the data five AOIs that corresponded to five criteria of the rubric were defined. According to Rayner (2009) there are two basic eye movements (saccades and fixation) and new pieces of information are acquired while eyes are fixated. Also, heat maps are based on fixations count and duration and the areas with more fixations are depicted in dense color. To examine which parts (areas of interest) of the analytic rubric the rater paid attention to, heat-maps were analyzed qualitatively to see how eye patterns and heat-map density changed during eight sessions of rater training.

5. Results

In what follows the results related to the heat-maps and the consequent feedback sessions will be elaborated.

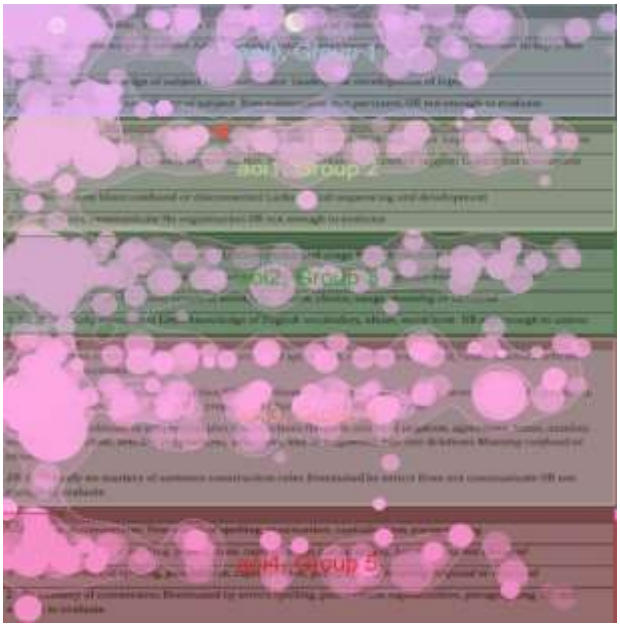

Figure 2. The Heat-map for the First Rater Trainig Sesison

According to the rater's eye record in the first attempt in essay scoring (Figure 2), it seems that the rater was mostly focusing on the allocated score for each criterion rather than the equivalent description.

Figure 3. The Heat-map for the Second Rater Trainig Sesison

Furthermore, the density of the heat-map is higher for four criteria (content, organization, language use, vocabulary) in comparison with the last one (mechanics) showing that the rater was paying less attention to this criterion. However, as rater training sessions continued (sessions 3,4,5) the rater tried to direct his attention to all criteria.

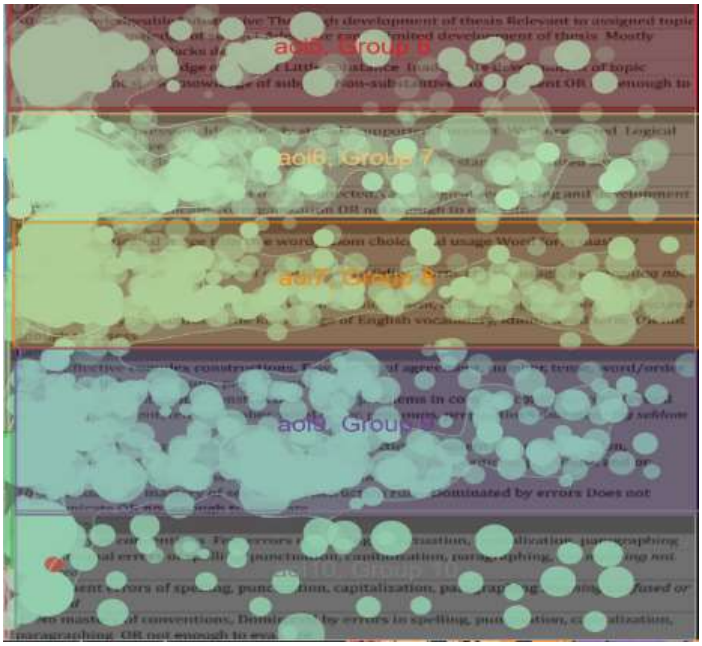

Figure 4. The Heat-map for the Third Rater Training Session

In session 3 (Figure 4), while the rater still focused more on the first four criteria, he allocated more attention to the last criteria (mechanics) in comparison with the first rater training session. Also, his eye movements showed that he read over criteria descriptions for organization, vocabulary and language use; which were neglected in previous efforts. As it is depicted in the fourth and fifth sessions of rater training, the rater has tried to focus on all criteria and the equivalent description for each to decide on the final score.

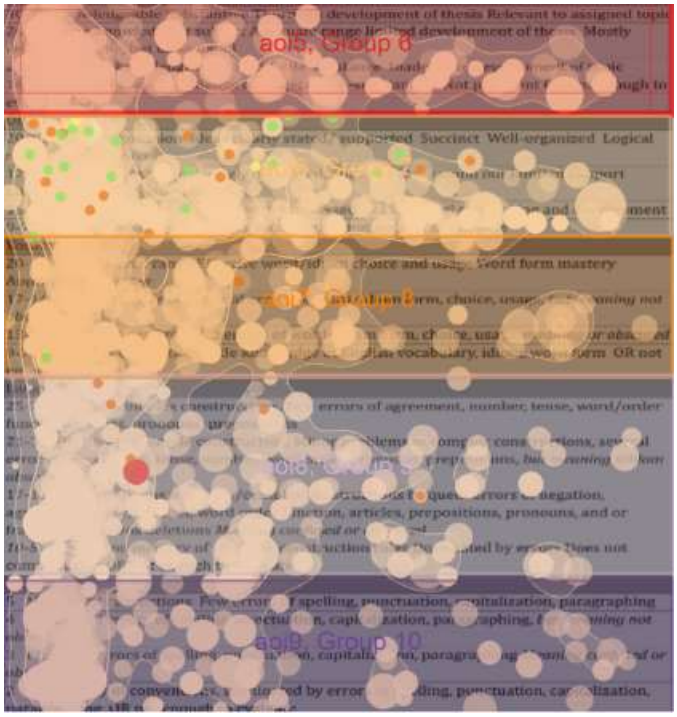

Figure 5. The Heat-map for the Fourth Rater Training Session

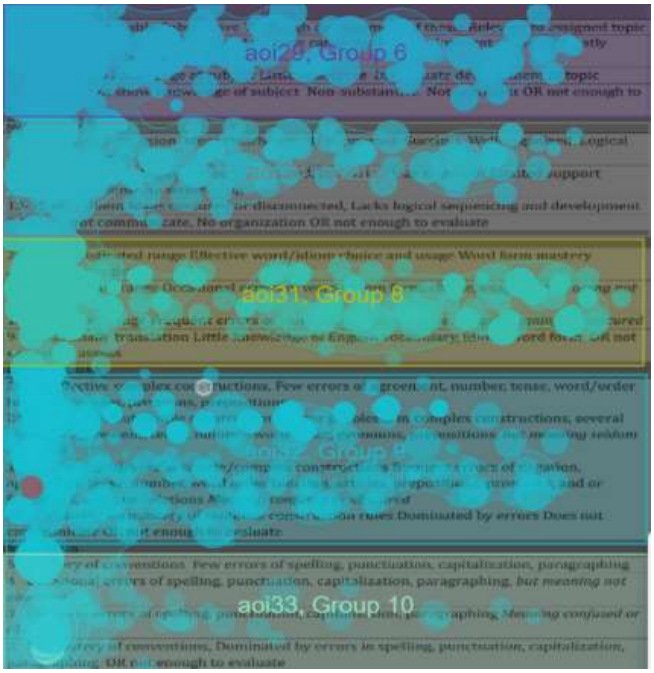

Figure 6. The Heat-map for the Fifth Rater Training Session

In the fourth and fifth sessions rater's eye patterns became more organized and all criteria were focused and considered similarly.

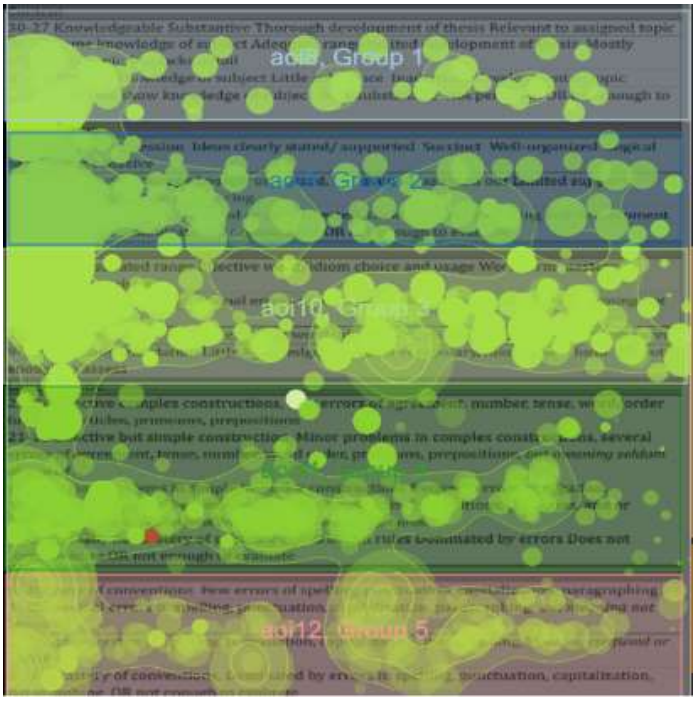

Figure 7. The Heat-map for the Last Rater Training Session

Finally, the heat map for the last session illustrates that the rater did focus on the last criteria (mechanics) which was mostly neglected in his previous attempts.

The current study was aimed at investigating the developmental process of rater training in a two-month training program through the use of eye tracking. The findings indicated that by the use of heat-maps and feedback sessions after each essay scoring the novice rater's eye patterns changed during sessions of rater training. For example, when in the fifth session he was asked to elaborate on the density of the heat map in mechanics section and to compare it with previous sessions, he asserted that:

> I am focusing more because I am more concentrated about spellings. And I try to, if I remember correctly, no, that's logical. We can say it was 5 minutes my first essay scoring I mean, because

> I just read over it once or maybe just twice for all the criteria, but you know, from session 3 or four on, I read them again and again, maybe 4 times, 5 times.

The heat-maps illustrated that through rater training sessions the effects of order and primacy in analytic rubrics is minimized. For the reason that the essence of analytic rubrics centers around the notion that raters should consider and focus upon all criteria equally (Knoch, 2009), primacy effect should be controlled and diminished. The study revealed that mechanics—a criterion initially receiving less focus from raters—emerged as a key factor in final essay scoring decisions. Moreover, following a few sessions of training, the rater tried to find the equivalent description that best defines the writer's performance rather than deciding on the score first, then looking for justifications in the descriptions provided. Previous rater training studies using think-aloud protocols have indicated that raters may mostly rely on their tentative score for each category after reading an essay rather than descriptors (Cumming et al., 2002; Lumley, 2002). However, in this study, the eye tracking heat-maps and the feedback session helped the rater to gradually modify such a behavior. For instance, in the 6th session in providing justifications for his scoring, the rater referred to the point that instead of assigning the score first, he preferred to read the criterion's description to make his final decision. As he said:

> I was considering of even 22 to 26. But like I if I hadn't like, read 17 to 21, I might have given it a 22. But when I read them again and again, when I read the third band score, I realized that the least score must be contributed to the essay.

These findings highlight the importance of rater training programs that emphasize the use of descriptors and encourage raters to reflect on their scoring processes. Training can help mitigate the reliance on tentative scores and promote a more consistent application of the rubric criteria (Erlam, 2013).

Previous rater training studies using eye tracking (Ballard, 2017; Winke & Lim, 2015) have reported the existence of primacy effect in the rating process. Findings of the current study also confirmed the existence of the

primacy effect in the early sessions of training; however, the current study unlike the previous studies which were only observational studies (depicting the existence of the primacy effect) was aimed at removing or minimizing the primacy effect and improving the rater behavior using eye-tracking output. This strategy aims to enhance the consistency and the objectivity of evaluations made using analytic rubrics, thereby improving overall assessment quality (Doğan, 2017; Gyamfi et al., 2022). The feedback sessions provided based on the results of eye tracking helped reduce the primacy effect and also helped raters to pay attention to all the scoring criteria and the relevant descriptors in assigning scores. This approach provides objective data about how raters visually engage with the rubric and the essays being evaluated, allowing for targeted feedback on their scoring behaviors (Ashraf et al., 2018). This is well documented by the rater, for example, in his following comment:

> Well, if I want to give you an example, there were times that I had, let's say, I was skipping towards the, I don't know, first paragraph or the last. And we could have noticed those skips by seeing that it possesses a light heat map instead of a heavy one, a dense one. And what we concluded [in the previous session] that the next time you should focus on other paragraphs as well. And as you can see, in my last essays, I've been more focusing more on the whole parts. I know the description of even, what are these called, umm criteria completely.

7. Conclusion and Implications

While previous eye-tracking studies have used eye tracking to observe and study raters' behavior, the findings of this study indicated that eye tracking could further be employed to help raters understand their rating problems and reduce the subjectivity in their rating. Eye tracking could be employed to study the developmental process of rating. Since the eye-tracking results are presented in terms of tangible heat-maps, rater trainers could easily employ the results of eye tracking after each rating to tell the raters how they rated the sample performance and where they need development. Raters could also

easily understand their own behavior by observing their heat maps. They could have a clear picture of which criteria they consider more in their rating, and whether they consider different descriptors while rating a sample of performance.

Finally, the study may have some limitations which should be considered in generalizing the findings. As a case study, it was limited to a single participant. This could potentially impact the applicability of the findings to a broader population. Additionally, the findings may have been affected by the individual characteristics of this participant.

References

Ary, D., Jacobs, L. C., Irvine, C. K. S., & Walker, D. (2018). *Introduction to research in education* (10th Ed.). Cengage Learning.

Ashraf, H., Sodergren, M. H., Merali, N., Mylonas, G., Singh, H., & Darzi, A. (2018). Eye-tracking technology in medical education: A systematic review. *Medical teacher*, *40*(1), 62-69.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Ballard, L. (2017). *The effects of primacy on rater cognition: An eye-tracking study*. Michigan State University.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, *7*(1), 54-74.

Bejar, I. I., Williamson, D. M., & Mislevy, R. J. (2006). Human scoring. *Automated scoring of complex tasks in computer-based testing*, 49-82.

Chen, K. T., Prouzeau, A., Langmead, J., Whitelock-Jones, R. T., Lawrence, L., Dwyer, T., ... & Goodwin, S. (2023, May). Gazealytics: A Unified and Flexible Visual Toolkit for Exploratory and Comparative Gaze

Analysis. In Proceedings of the 2023 Symposium on Eye Tracking Research and Applications (pp. 1-7). Preprint available at arXiv:2303.17202.

Conklin, K. & Pellicer-Sánchez, A. (2016). Using eye-tracking in applied linguistics and second language acquisition research. *Second Language Research, 32*(3), 453-467.

Cumming, A. (1990). Expertise in evaluating second-language compositions. *Language Testing*, *7*(1), 31-51.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal*, *86*(1), 67-96.

DeRemer, M. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing,* 5, 7–29.

Deygers, B., & Van Gorp, K. (2015). Determining the scoring validity of a co-constructed CEFR-based rating scale. *Language Testing*, *32*(4), 521-541.

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). Factors in judgments of writing ability. *ETS Research Bulletin Series*, *1961*(2), i-93.

Dogan, C. D., & Uluman, M. (2017). A Comparison of Rubrics and Graded Category Rating Scales with Various Methods Regarding Raters' Reliability. *Educational sciences: Theory and practice*, *17*(2), 631-651.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*(3), 197–221.

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Peter Lang.

Eckstein, G., Casper, R., Chan, J., & Blackwell, L. (2018). Assessment of L2 student writing: Does teacher disciplinary background matter? *Journal of Writing Research*, *10*(1), 1-23.

Elder, C., Knoch, U., Barkhuizen, G., & Von Randow, J. (2005). Individual feedback to enhance rater training: Does it work?. *Language Assessment Quarterly: An International Journal*, *2*(3), 175-196.

Elder, C., Barkhuizen, G., Knoch, U., & Von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, *24*(1), 37-64.

Engelhard Jr, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge.

Erguvan, I. D., & DÜNYA, B. A. (2021). Gathering evidence on e-rubrics: Perspectives and many facet Rasch analysis of rating behavior. *International Journal of Assessment Tools in Education*, *8*(2), 454-474.

Erlam, R., von Randow, J., & Read, J. (2013). Investigating an online rater training program: product and process. *Papers in Language Testing and Assessment*, *2*(1), 1-29.

Godfroid, A. (2019). Investigating instructed second language acquisition using L2 learners' eye-tracking data. In *The Routledge handbook of second language research in classroom learning* (pp. 44-57). Routledge.

Godfroid, A., & Spino, L. A. (2015). Reconceptualizing reactivity of think-alouds and eye tracking: Absence of evidence is not evidence of absence. *Language Learning*, *65*(4), 896-928.

Godfroid, A., Winke, P., & Conklin, K. (2020). Exploring the depths of second language processing with eye tracking: An introduction. *Second Language Research*, *36*(3), 243-255.

Gyamfi, G., Hanna, B. E., & Khosravi, H. (2022). The effects of rubrics on evaluative judgement: a randomised controlled experiment. *Assessment & Evaluation in Higher Education*, *47*(1), 126-143.

Hamp-Lyons, L. (2007). Worrying about rating. *Assessing Writing*, *1*(12), 1-9.

Harsch, C., & Martin, G. (2012). Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing*, *17*(4), 228-250.

Jacobs, H., Zinkgraf, S., Wormuth, D., Hartfiel, V., & Hughey, J. (1981). *Testing ESL composition: A practical approach. Rowley*. Newbury House.

Janssen, G., Meier, V., & Trace, J. (2015). Building a better rubric: Mixed methods rubric revision. *Assessing writing*, *26*, 51-66.

Jin, K. Y., & Eckes, T. (2022). Detecting differential rater functioning in severity and centrality: The dual DRF facets model. *Educational and Psychological Measurement*, *82*(4), 757-781.

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, *26*(4), 485-505.

King, A. J., Bol, N., Cummins, R. G., & John, K. K. (2019). Improving visual behavior research in communication science: An overview, review, and reporting recommendations for using eye-tracking methods. *Communication Methods and Measures*, *13*(3), 149-177.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, *26*(2), 275-304.

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior—a longitudinal study. *Language Testing*, *28*(2), 179-200.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing writing*, *12*(1), 26-43.

Li, Y., Wei, C., & Ma, T. (2019). Towards explaining the regularization effect of initial large learning rate in training neural networks. *Advances in Neural Information Processing Systems*, *3*(2), 1-49.

Linacre, J. M. (2004). Optimizing rating scale effectiveness. In E. V. Smith & R.M. Smith (Eds.), Introduction to Rasch measurement (pp. 257–578). JAM Press.

Low, A. R. L., & Aryadoust, V. (2021). Investigating test-taking strategies in listening assessment: A comparative study of eye-tracking and self-report questionnaires. *International Journal of Listening*, *35*(1), 1-20.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, *19*(3), 246-276.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. P. Lang.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language testing*, *12*(1), 54-71.

Luoma, S. (2004). *Assessing speaking*. Cambridge University Press.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of applied measurement*, *4*(4), 386-422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of applied measurement*, *5*(2), 189-227.

Rayner, K. (1978). Eye movements in reading and information processing. *Psychological bulletin*, *85*(3), 618.

Rayner, K. (2009). Eye movements in reading: Models and data. *Journal of eye movement research*, *2*(5), 1.

Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. *Language testing*, *25*(4), 553-581.

Saslow, J., & Ascher, A. (2015). Top notch (3rd ed.). Pearson Education.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25*(4), 465-493.

Shin, Y. S. (2009). A FACETS analysis of rater characteristics and rater bias in measuring L2 writing performance. *English Language & Literature Teaching*, *16*(1), 123-142.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal*, *76*(1), 27-33.

Stewart, A. J., Pickering, M. J., & Sturt, P. (2004). Using eye movements during reading as an implicit measure of the acceptability of brand extensions. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *18*(6), 697-709.

Suto, I. (2012). A critical review of some qualitative research methods used to explore rater cognition. *Educational Measurement: Issues and Practice*, *31*(3), 21-30.

Vaughan. C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Ed.) *Assessing second language writing in academic contexts*, 111-125.

Wang, J., & Engelhard Jr, G. (2019). Exploring the impersonal judgments and personal preferences of raters in rater-mediated assessments with unfolding models. *Educational and Psychological Measurement*, *79*(4), 773-795.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*(2), 197-223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263-287.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, *6*(2), 145-178.

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

Wind, S. A. (2019a). A nonparametric procedure for exploring differences in rating quality across test-taker subgroups in rater-mediated writing assessments. *Language Testing*, *36*(4), 595-616.

Wind, S. A. (2019b). Examining the impacts of rater effects in performance assessments. *Applied Psychological Measurement*, *43*(2), 159-171.

Wind, S. A., & Peterson, M. E. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, *35*(2), 161-192.

Winke, P., & Brunfaut, T. (Eds.). (2021). *The Routledge handbook of second language acquisition and language testing*. Routledge.

Winke, P., & Lim, H. (2015). ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study. *Assessing Writing*, *25*, 38-54.

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing*, *4*(1), 83-106.

Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *Language Testing, 31*(4), 501–527.

Youn, S. J. (2018). Rater variability across examinees and rating criteria in paired speaking assessment. *Papers in Language Testing and Assessment*, *7*(1), 32-60.