Scoring Behavior of Language Instructors(*)

Under the Supervision of
Amira Agameya

Marwa Mohamed Essam Saifalnasr Baza

The American University

Abstract

This experimental study made use of different qualitative and quantitative analyses to uncover instructors' scoring behaviors of Reading-to-Write tasks in Egypt. Egyptian instructors working in the field of higher education in Egypt were requested to score Reading-to-Write essays following the Think Aloud Protocol, assign scores to the different writing features on the analytic rubric, and be interviewed. Based on the quantitative and qualitative data, it was clear that instructors paid most of their attention to judgment strategies, and most specifically to rhetorical/ ideational aspects with little attention to language focus. The central tendency in grading was mostly observed in the scoring process. All raters followed a pattern of having a mental image representation of scores once they started scoring and either confirmed this or changed it once reading was completed. Instructors did not follow the order of the analytic rubric used and instead assigned scores to the features that stood out the most because they were either too good or too bad. No clear pattern of severity or leniency was observed due to the limited number of participants although it was noticed that the instructor who was mostly severe almost always made negative comments when scoring, whereas the rater who was mostly lenient was the one who was generally sympathetic with the students when scoring.

Key Words: Reading-to-Write Tasks, Scoring Behavior, Judgement Strategies, Central Tendency, Analytic Rubrics

(*) Scoring Behavior of Language Instructors, Vol.11, Issue No.2, April 2022, pp. 39-61.

الملخص

استندت هذه الدراسة التجريبية على التحليلات النوعية والكمية المختلفة للكشف عن سلوكيات المدربين في تقييم مهام القراءة للكتابة في مصر. طُلب إلى المدرسين المصريين العاملين في مجال التعليم العالي في مصر تقييم مقالات القراءة وللكتابة مستخدمين بروتوكول Think Aloud وهو التفكير بصوت، وتعيين درجات تبعًا لنموذج التحليل (الروبريك التحليلي)، وإجراء مقابلات معهم. استنادًا إلى البيانات الكمية والنوعية، كان من الواضح أن المدرسين قد أولوا معظم اهتمامهم لاستراتيجيات الحكم، وبشكل أكثر تحديدًا للجوانب البلاغية/ الفكرية مع القليل من الاهتمام للتركيز اللغوي. تم ملاحظة الاتجاه المركزي في الدرجات. اتبع جميع المقيّمين نمطًا لتمثيل الصورة الذهنية للعلامات بمجرد أن بدأوا التسجيل وسجلوا تأكيدًا لذلك أو غيروه بمجرد اكتمال القراءة. لم يتبع المدربون ترتيب نموذج التقييم التحليلي المستخدم وبدلًا من ذلك قاموا بتعيين درجات للميزات التي كانت أكثر بروزًا؛ لأنها إما جيدة جدًّا أو سيئة للغاية. لم يلاحظ أي نمط واضح للشدة أو التساهل بسبب العدد المحدود من المشاركين على الرغم من أنه لوحظ أن المعلم الذي كان شديدًا أبدى تعليقات سلبية عند التسجيل، في حين أن المقيم الذي كان متساهلًا هو الشخص الذي كان متعاطفًا بشكل عام مع الطلاب عند التقييم.

الكلمات المفتاحية:

مهام القراءة للكتابة- السلوك التقييمي- استراتيجية الحكم- النزعة المركزية في التقييم- النماذج التحليلية (الروبريك التحليلي).

Integrated writing tasks have received considerable attention in the literature lately, especially in the field of language testing and assessment. Integrated Writing Tasks can be defined in many ways, the most common of all is the one introduced and summarized by Knoch and Sitajalabhorn (2013), where they stated that all definitions in the literature involved the idea of having a source that included a variety of useful ideas or thought-provoking issues that the examinees could make use of in writing. The overall process of introducing a source-be it in the form of a listening prompt, a reading prompt, both listening and reading prompts, or a graph- and asking students to write about it by explaining,

summarizing or arguing, for example, led scholars, instructors, and testers to categorize the task as an Integrated Writing Task.

Examples of such tasks used in language tests include summary writing, writing about graphs and charts as in Task 1 IELTS or reading-to-write along with listening prompts as in Test of English as a Foreign Language Intenrt-Based Test(TOEFL iBT). Clearly, these are different forms of writing that involve different cognitive processes, critical thinking abilities, as well as different proficiency levels of reading and/or writing.

Upon thinking about rater perceptions and attitudes, educators need to worry about what raters deemed acceptable and what they deemed unacceptable, and how different raters approached writing tasks and rubrics. The values raters placed on the different criteria is one reason why score variance was observed in various examination settings. Eckes (2008) had this to say about the different types of raters.

> For example, raters may differ (a) in the degree to which they comply with the scoring rubric, (b)in the way they interpret criteria employed in operational scoring sessions, (c) in the degree of severity or leniency exhibited when scoring examinee performance, (d) in the understanding and use of rating scale categories, or (e) in the degree to which their ratings are consistent across examinees, scoring criteria, and performance tasks. (Eckes, 2008, p. 156)

Eckes hypothesized that there were different rater types based on their 'reading-style' and foci they paid special attention to. To prove his hypothesis, he asked 65 raters of TestdaF (test of German as a foreign language called ***Test Deutsch als Fremdsprache***) to rate different criteria of a scoring rubric as per the level of their importance. The outcome was that raters varied in the proportions of weight given to different criteria. Accordingly, Eckes (2008) concluded that there were six types of raters who had criterion-based foci; that is, different raters had different weight

placed on different criteria based on how important they perceived each criterion to be. Four types of raters were dominant, and these were: The Correctness, The Syntax, The Fluency, and The Structure types. The other two types were measured based on the criteria raters did not focus on, and these were: The Non-Fluency and The Non- Argumentation types.

In a follow up study, Eckes (2012) hypothesized that 'rater cognition' would shape 'rater behavior'. According to the evaluation of 18 raters in his study, Eckes found that the severity/ leniency continuum was attributed to raters' beliefs. Accordingly, the more importance raters placed on a specific criterion, the more severe they were in grading it. Likewise, the less value the raters placed on some criteria, the more lenient they were in grading them. Eckes then suggested that the issue of rater cognition and rater bias should be taken into consideration when holding norming sessions to ensure that raters were aware of their biases to try to work around them and achieve better consistency among raters.

This issue of rater bias was also clear in Schaefer's (2008) study. In his study, 40 native speakers rated 40 EFL essays of TWE of female Japanese students where he proved that rater bias did exist. Among the interesting results of his study was the fact that the higher the student proficiency was, the more severe the rater was, and the lower proficiency the student had, the more lenient the rater was. Another pattern of rating also emerged in his study, and that was seen as a 'compensatory' strategy. That meant that when some raters showed severe rating on some criteria, they proved quite lenient with the other set of criteria. For example, when a rater was severe on the 'content and/ or organization' criterion, he was lenient on the 'language and/or mechanics' criterion. Some reversed pattens also appeared. As with Eckes (2008, 2012), Schaefer (2008) suggested that these accounts needed to be addressed in norming sessions to increase the level of 'rater self-awareness' and achieve accuracy.

Rater inconsistencies were also discussed in other studies. One of the reasons provided was referred to as 'ego involvement', which

reflected a level of subjectivity on the part of raters (Wiseman, 2012). In her study, Wiseman indicated that raters tended to be more lenient when they got rather personal with the essay writer. When raters tried to reword students' essay to understand, got impressed by a writing style or some general knowledge, they gave higher evaluations through 'self-monitoring'. Wiseman referred to that as the 'mitigating effect' where raters' background and expectations interfered with how strict they were and how focused on the descriptors they were, too. Of course, this was quite evident with the holistic scoring more than with the analytical scoring although it was evident in both.

Looking at scoring behavior from another angle, reference needs to be made to Huot (1993, as cited in Ohta, 2018) where the difference between experienced and inexperienced raters when scoring L1 writing was pointed out. Experienced raters were found to make their overall, final evaluations after reading the whole essay and even showed more personal engagement with the writers**.** Scott and Bruce (1995, as cited in Baker, 2012) also came up with an interesting 'inventory' for different scoring behaviors that they referred to as decision making styles, which were: 'rational, dependent, intuitive, avoidant, and spontaneous'

In her study, Baker (2012) concluded that the most common scoring behaviors exhibited by the six raters in her study after analyzing some write-alouds (raters wrote down the thoughts and feelings that guided their graading). Such write-alouds were done while grading a reading-to-write exam (EETC-a teacher certification exam in Quebec) and filling out a questionnaire about decision-making styles were the rational and intuitive styles. Likewise, the 'spontaneous', the 'dependent', and the 'avoidant' were also evident, the latter exhibiting some 'avoidance techniques' by refraining from giving lowest or highest scales on the rubric; that is, they preferred to give the middle range of scores. In relation to these decision-making styles, reference could be made to the four approaches exhibited by raters as observed by Milanovic, Saville, and Shuhong (1996), and these were the 'provisional

marker' (who makes scoring decisions at an early stage of reading), the 'principled two-scan/reader' (who always reads twice before deciding on the score), the 'pragmatic two-scan/reader' (who reads twice when hesitant or unsure), and the 'read-through' reader (who grades based on intuition or impression)- as cited in Ohta (2018).

Rater and rurbic-related issues were also observed with L2 Integrated Writing. One of the earliest studies that tried to look into raters' scoring behavior in L2 was Cumming et. al.'s (2001) three-stage study where she uncovered the scoring behavior of 17 assessors-both native and non-native speakers-through TAPs scoring a number of independent and integrated writing tasks (five different types of L-W and R-W) for the new TOEFL 2000 prototype. The raters looked at 35 scoring strategies employed by assessors. Those 35 strategies detailed two main scoring behaviors: a) Judgement, implying making evaluation decisions about performance and b) Interpretation, implying making sense and understanding the writer's position. Such Judgement and Interpretation strategies were clear through three main scoring foci: a) 'self-monitoring focus', implying the attempts to understand the overall task and students' overall task completion; b) rhetorical/ ideational focus', implying the attempts to understand the students' content, style, rhetoric, cohesion, and organizational patterns; and c) 'language focus', implying attempts to attend to language proficiency aspects, including grammar, spelling, and vocabulary. In this study, it was noticed that the judgment behaviors exhibited were more than the interpretation behaviors shown by raters although both received considerable attention. Most importantly, the raters assessing the integrated tasks paid more attention to rhetoric and content than they did to language focus. It was also observed that native raters focused more on rhetoric, whereas non-native raters were more focused on language patterns.

Following Cumming, Kantor, & Powers (2001), Gebril and Plakans (2014) adapted their framework of 35 strategies and used 31 strategies only. They still followed the same strategies of a) judgement and b) interpretation. They also followed the same pattern of three foci:

a) self-monitoring, b) rhetorical/ ideational, and c) language. In their attempt to unveil the rating processes and scoring behavior of raters, Gebril and Plakans conducted some inductive analyses on TAPs and interviews with two experienced raters: one native and one non-native. The two raters scored 145 EFL reading-to-write essays of students in a Middle eastern University. The results of their analyses showed that raters had some challenges when it came to assessing source use; these included how to locate the information in the sources and how to judge the success of integration on the level of quality and citation mechanics. These were all matters pertaining to evaluating source use-an element of much importance on any rubric of integrated tasks. Other than attending to issues related to 'source use', all raters in the study seemed to exhibit more 'judgement strategies' than 'interpretation strategies'.

Shin and Ewert (2015) also noted that "to date, raters' behaviors in terms of their severity or leniency, and the effects of these behaviors on score reliability of the changing number of raters in different analytic scoring domains of the RTW [Reading-To-Write] task have not yet been sufficiently investigated" (p.262). This shows the need to look into issues pertaining to rater behaviors in the area of Reading-Based-Writing Tasks. Added to that, few studies handled issues of rater behavior, challenges, and cognitive processes when it came to Integrated Writing Tasks, the latest of which is that of Gebril and Plakans (2014) in which they explored rating behavior and challenges when it came to assessing Integrated Writing Tasks. However, the data provided was based on the information (derived through grading, Think Aloud Protocols (TAPs), and interviews) of two raters.

One of the primary tools that has been used in research to uncover the rating processes and strategies employed by raters when scoring is the Think Aloud Protocols commonly referred to as TAPs or TA. These are verbal accounts where raters speak out loud to explain their thinking process while reading essays and assigning scores. According to Barkaoui (2011), the process could be summarized as follows, ' Raters' verbalizations of their thoughts are recorded, transcribed, and then

analyzed to identify the decision-making processes that raters employ and the aspects of writing they attend to when marking essays' (p.51).

However, as much as the Think Aloud technique was regarded a good way to uncover what went on in the minds of raters in a very specific way (Barkaoui, 2011), it was also criticized for the pressure it added on raters while scoring. Other than having to attend to the careful reading and judgement of essays, raters had to conscientiously record their thoughts every step in the way until a final scoring decision was reached. Accordingly, the TAPs added one more level of difficulty to the scoring process which was already complex in nature (Winkie & Lim, 2015).

In that light, some criticism had been waged against the TAPs based on their 'reactivity' (i.e. it was difficult to fully and clearly verbalize one's thoughts) and 'veridicality' (i.e. TAPs could never display the full picture of the cognitive processes of raters) (Barkaoui, 2010; Barkaoui, 2011; Winkie & Lim, 2015). That is why Barkaoui (2011) recommended using TAPs with other means of data collection to ensure capturing the full range of information required (i.e. interviews, questionnaires, etc.).

Based on the above and since to the best knowledge of the researcher, no research had been conducted, especially in Egyptian English-medium universities, to focus on instructor-rater scoring behavior when scoring Reading-to-Write Tasks using analytic rubrics, this study hoped to contribute to the field by filling this gap in the literature. The study aimed to examine the scoring behavior of instructor-raters when assessing a Reading-Based-Writing Task. The research question that was investigated was:

What are the raters' scoring behaviors in Reading-to-Write tasks?

In order to provide a clear answer to this question four instructors were requested to volunteer to help the researcher. Those who agreed to help out were four female, Egyptian instructors with a wide range of teaching experience. Their experience ranged from 20 years to 34 years.

Qualitative data as well as quantitative data was used. In terms of the qualitative data, they were received through 60 Think Alouds obtained through 15 Reading-to-Write essays that were scored by four instructors. Hence, each instructor provided 15 Think Alouds while grading the essays, so the total was 60 Think Alouds. These were recorded using *Audacity* Software- a free online software that instructors downloaded on their PCs and laptops.

All instructors were requested to attend a training session to know how to approach the Think Alouds. Two researchers attended a training together, whereas the other two sat with the researcher individually due to conflicting time-tables. In these sessions, they were exposed to the process of how to use *Audacity* (a free recording software) to record their thoughts for the Think Alouds. Following Gebril and Plakans (2014) model of training, I explained to them how to clearly record their thoughts in a loud voice. To do that, a sample was provided to them, followed by an example led by myself. After that, I asked them to verbalize their thoughts while answering some addition and subtraction equations. They were also requested to verbalize their thoughts while grading a paragraph or two using the Think Aloud Protocol to see how they would perform.

Once the four instructor-raters were all comfortable with the idea of recording and knew how to do it, they were given the 15 essays to grade at their own pace. It is important to note that the instructors were requested to record the grades they assigned in table format provided to them. This table format followed the order of the writing features of the analytical rubric they used: integration, content, organization, vocabulary, and grammar- a 5 Likert scale multi-trait rubric.

Another source of qualitative data was obtained through interviews held with the same four instructors after completing the Think Alouds and scoring of essays (See Appendix A). The four raters who took part in both the interviews and the Think Alouds were all female, experienced raters with experience in teaching writing for a minimum of 20 years and a maximum of 34 years.

As for the quantitative data, these were obtained through the scores assigned by the four raters to the 15 essays following the analytic rubric they had been using for years. Table 1 provides as example of the table used for assigning the scores.

Table 1

Table of Assigning Grades on a scale from 1-5 (1= very poor- 5= excellent)

	Integration of Sources (20%)	**Content 20%**	**Organization (20%)**	**Vocabulary (20%)**	**Grammar (20%**	**Total (100%)**
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						

For the analysis, the following procedures were followed:

a) aggregated frequencies and percentages from the Think Alouds

b) inductive analyses of the Think Alouds and interviews.

c) quantitative analysis of the scores assigned to the total of 60 Think Alouds

Results

The four raters in the interviews said that they graded following the analytic rubric although two of them said that they had some aspects to think about in relation to this point before deciding on a final score to assign. Rater 1 said, *"I mostly stick to the analytic rubric, but I also look at a paper holistically before I make a final decision. I believe both should complement each other to reach an accurate evaluation of a paper."* Hence, while Rater 1 said she followed the holistic approach in tandem with the analytic rubric, Rater 3 said that she tended to follow the analytic scale while giving attention to her '*gut feeling*' and what was expected of the students at the following level. Accordingly, as all four raters put it in their interviews, they had mental analysis of the essays they were reading and had an estimate of what the students would score; this initial score was either confirmed or changed when they completed reading the essay at hand.

Rater 3 said she might not read till the very end. She specifically had this to say about this point when interviewed: "… *I think I skip the quotes, and I know these are copied, so I don't need to read them, and many times I have determined the grades before the conclusion. I know that is wrong, but I would do that. But I would never skip the introduction or the body paragraphs*." This was confirmed in her Think Alouds when commenting on paper 10: "*I've already decided the grade before reading the conclusion anyway."* The three other raters expressed the necessity of completing the whole essay before assigning a score as '*some students might surprise you'* with an unexpected performance towards the end of the essay, as Rater # 1 put it.

When it came to the Think Alouds, a set of analyses were conducted with a colleague (who acted as coder 1) following Gebril and Plakans (2014) who, in turn, employed a model created by Cumming et

al. (2001). This model depended on a set of frequency counts that specifically looked at 31 strategies employed by the raters who used the TAPs when grading the Reading-to-Write tasks. This model was divided into three main categories. The three categories were 'a) self-monitoring, b) rhetorical/ ideational, and c) language focus'. 'Self-monitoring' referred to the behaviour exhibited by the different raters to check their understanding of the prompt, rubric, or task; 'rhetorical/ ideational' had to do with how far the raters focused on the ideas raised and the writing style used by the students; and 'language focus' was concerned with the different foci related to language use as in grammar, punctuation, fluency, spelling, and the overall control shown by the students in this aspect. Each of these three categories was subdivided into two strategies: a) *Interpretation Strategies* (the raters' attempts at understanding the different elements they encountered in order to be able to grade the essays fairly) and b) *Judgement Strategies* (the raters' remarks when making decisions, such as evaluating the essays or assigning scores).

After careful analyses of the Think Alouds where frequency counts of each strategy was accounted for, it became evident that all four raters mostly exhibited judgement strategies (73%) as opposed to interpretation Strategies (27%) as see in Table 2.

Table 2

Overall Percentages of Strategies Employed by all Four Raters

Strategies	**Average of Aggregated Percentages of Four Raters**
Interpretation Strategies	27%
Judgement Strategies	73%

Note. The average of percentages was rounded.

Also, it was noticed that all rates reflected high use of rhetorical/ ideational strategies (55.6%) followed by language focus strategies (37.8%), while self-monitoring strategies were the least reflected standing at (6.6%) as detailed in Table 3.

Table 3

Comparison of the Three Scoring Behaviors Among Four Raters

	Self-Monitoring	Rhetorical/ Ideational	Language Focus
Rater 1	6.155%	55.38%	38.455%
Rater 2	5.015%	47.24%	47.74%
Rater 3	5.64%	56.49%	37.66%
Rater 4	9.385%	63.395%	27.215%
Average	**6.6%**	**55.6%**	**37.8%**

Note. The average was rounded to one decimal point

If a comparison across raters were to take place, it would be noticed that Raters 1, 3, and 4 paid most of their attention to the rhetorical/ ideational decision-making behaviors with (55.38%), (56.49%), and (63.38%), respectively as clear in Table 3 above. They were all close in numbers except for the fact that Rater 4 had the highest percentage in this area. Perhaps this could be explained in light of the Think Alouds of Rater 4. In fact, it was noticed that Rater 4 did not make as many comments as the other raters did. It was almost always a summary given at the end of each paragraph summarizing it and evaluating the performance at every stage. Perhaps this could explain the highest percentage of 63.38% dedicated to the rhetorical/ ideational behaviour and the lowest percentage of 27.22% dedicated to language use by the same rater (Rater 4). Rater 2, in contrast, showed a strikingly different behavior, where equal attention was given to both rhetorical/ ideational and language focus standing at (47.24%) and (47.74%), respectively. Hence, it was clear that Rater 2 attended to both aspects equally and did not value one more than the other.

Also-as clear in Table 4, Rater 2 was the one who scored the highest percentage in exhibiting judgement making behaviour (76.5%) compared to interpretation (23.5%). Likewise, all other raters showed more judgement behavior than interpretation behavior. They clearly spent

most of the time assessing content, development, integration, source use, sentence structure, and how successful the students were in general in achieving the target being assessed.

Table 4

Decision Making Behaviors of Raters

Strategies	**Average count of Coders**
Rater 1 Interpretation	26.50%
Rater 1 Judgement	74.50%
Rater 2 Interpretation	23.50%
Rater 2 Judgement	76.50%
Rater 3 Interpretation	28%
Rater 3 Judgement	72.50%
Rater 4 Interpretation	31%
Rater 4 Judgement	70%

Trying to uncover the details related to raters' characteristics in terms of leniency and severity, data was extracted from the qualitative data from the interviews and quantitative data from the grades assigned when doing the Think Alouds of the 15 essays. Based on the grades assigned to students to the 15 students by the four raters, it was clear that Raters 1 and 2 were very close to one another, and it was also clear that they were almost always in the middle range of scores. On a scale from 1-5 on the rubric, they were always in the range from 3-4 as clear in Table 4. When asked about this central tendency in interviews, all raters said that many raters tended to do that for different reasons. Rater 3 said:

> The rubric is skimpy allowing for us to go to the central to be on the safe side. I also think I do that too many times especially when I am exhausted and have corrected a whole batch and still

> am required to correct another batch, all in one sitting. I think going to the central grade makes me feel that the other reader will determine where this essay falls, instead of me making that decision while I am exhausted and over worked.

Rater 1 agreed with rater 3 when she said:

> I think many raters do that to be on the safe side or not to be labeled as tough or lenient or inconsistent. I sometimes do that when I am not very sure about a paper or when I get confused how to grade it on one or more criteria. I can also say that I rarely go to extremes either way.

Rater 4 said the same thing as Rater 1 but added: '*Most teachers want to be on the safe side and want to avoid being judged as 'generous' or 'tough'*. However, Rater 2, seemed to disagree as she said, '*Most teachers lean towards in order to avoid re-reads and give weak students the benefit of the doubt. I don't.*' It was clear then from the words of interviewees that the central tendency was often practiced, especially when there was a lot of pressure placed on instructors to finish a lot of essays within a short period of time. When graders were tired, they would rather lean towards a middle grade leaving the final judgment to their co-grader. Most importantly, it was evident that many graders were worried about how others would judge them, so they would resort to the middle score on the rubric to avoid blame or attack.

This central tendency was not as clear though in the grades assigned by Raters 3 or 4 in their TAPs. In fact, Rater 3 was always assigning lower scores than others. The average of scores assigned was around 2.9 in most cases, while Rater 4 was almost always on the high end of the scale awarding a lot of (4)s on the different criteria. Her average was around 3.5 (Table 5). Perhaps Rater 3 did not go for the central set of scores this time as she sometimes did as she mentioned in her interviews since she knew there were no second readers and no actual

exam; in other words, the situation here was not high stakes. In fact, the scores she assigned were more in line with describing herself as tough grader although she said this was how other raters viewed her- not how she viewed herself. When asked to place herself on a continuum from 1-10 (from the most lenient to the most severe), she gave herself an 8; thus, leaning towards harsh. As for Raters 1, 2, and 4, they gave themselves a 6, which meant that they saw themselves as average graders. Rater 1 specifically said that she did not see herself as lenient or severe; she actually saw herself as 'accurate', whereas Rater 4 preferred to describe herself as 'fair'- not lenient or severe.

Table 5

Average scores Assigned by Raters in TAPs

Rater #	Av Integration	Av Content	Av Organization	Av Vocab	Av Grammar	Av Total
Rater 1	3.37	3.57	3.60	3.40	3.23	**68.87**
Rater 2	3.23	3.27	3.37	3.30	3.50	**66.73**
Rater 3	2.80	2.83	2.90	3.00	3.20	**58.83**
Rater 4	3.40	3.50	3.43	3.53	3.57	**69.63**

Discussion

According to the qualitative results, raters exhibited much more judgment strategies than interpretation strategies when rating this Reading-to-Write task, and this coincide**d** with Cumming et al. (2001) and Gebril and Plakans (2014) who concluded that their participants reflected more judgement strategies than interpretation ones. This could be explained considering scoring Reading-to-Write tasks using the analytic rubric. Raters had to make detailed decisions about different writing features. Raters were requested to think aloud while scoring and assigning scores based on the analytic rubric provided to them. Accordingly, the judgment behavior was expectedly dominant in this study. What was surprising in the results was the minimal self-monitoring focus practiced by the four raters who took part in the study-

the point which contradicted the findings of Barkaoui (2010) who concluded that raters using analytic rubrics would reflect both judgment and self-monitoring behavior.

This difference could be explained given the raters' background in this study, all of whom were experienced raters who spent more than twenty years teaching language courses, at least five of which included teaching integrated writing. The raters also used the analytic rubric that they had been using for years in their department. Accordingly, they did not have to reread parts of the rubric or adjust their status. In agreement with Wang (2014), the experienced raters mentioned in their interviews that the rubric was almost memorized and 'internalized'; they only resorted to it with complicated essays that were rather difficult to score and mainly to reconcile their initial judgement with the rubric-as suggested by Eckes (2008).

Most of the judgement strategies exhibited were dedicated to the rhetorical/ ideational aspects of the essays. Raters focused on content and integration. They displayed a lot of focus on the language, too. They made comments pertaining to punctuation, spelling, and all aspects of language, especially syntax. However, this language focus was lesser than the ones reflected in the area of rhetoric and content. Perhaps this was due to the nature of the task that demanded careful consideration to the ideas and the source being read to be able to judge the proficiency level-an important difference between this task type and independent tasks.

In line with previous research, too, was how raters explained their rating procedures. Rater 1 though said that she would sometimes skip reading the obvious (i.e. quotations), whereas Rater 3 said that she often did not read the conclusion since the students' proficiency usually became clear at an early stage of reading. Rater 3 could be described in this case as the 'provisional rater' who would decide on a score early in the reading process as described by Milanovic, Saville, and Shuhong (1996)- as cited in Ohta (2018). In relation to this was how all raters said that they would always start off with a score in their mind (i.e. mental

representation) which is either confirmed or changed throughout the reading. Lumley (2005) referred to this and called it 'the first reading stage' where decisions about scores take place. According to Lumley, this was followed by two other stages in which scores were assigned/ justified and spelled out or written.

Although all raters said that they strictly followed the rubics, two raters said that they followed their own impressions in tandem with the analytic rubric. Rater 1, for example, said that she used a holistic approach to complement her scores on the analytic scale. This approach was similar to that of Lumley's (2002) who said that raters often tried to find some middle ground between their own impressions and the rubric they were using. Likewise, Rater 3 said that she considered two things when assigning a score: a) her 'gut feeling' and b) the following level that the students should be placed in after the scoring is done. Accordingly, as suggested by Smith (2000, as cited in Barkaoui, 2010) some aspects raters attended to in assessment were not always listed in rubrics, and this definitely needed careful analysis to ensure good written rubrics and fair assessment that was consistent across all raters.

In the interviews, the four raters said that they would start assigning scores to the feature that stood out or that there was no confusion about while reading-either because it was too good or too bad. Hence, the order of the rubric was not of much value in this area. Knowing this, a question needs to be raised: Do raters start with assigning a score to the feature that they mostly focus on? Do they start with assigning a score to the feature that designates their reading style or rater style? This is a point which needs to be carefully looked at if rater self-awareness-or rater-bias- was to be discussed in norming sessions.

Among scoring-related issues was how raters tended to resort to the central tendency pattern in scoring- a point that was raised in literature by Engelhard (1994, as cited in Wang (2014, and Wiseman, 2012). This issue was also referred to as the 'avoidance' strategy as suggested by Baker (2012). While investigating this issue of central tendency in interviews, raters reported that they sometimes avoided

giving very low scores or very high scores, except for one rater who said she would stick to the rubric and its descriptors regardless of where it would place students. Raters explained that when essays were difficult to score, they would go for the middle range to a) give the students' the benefit of the doubt or b) to place the responsibility on the second grader when they felt confused or exhausted. This eased the burden placed on their shoulders if they suspected making wrong judgements. This shows how conscientious raters were, but it also reflected how exhausted they often were, especially when required to score many batches within a limited time frame as explained in interviews. Another reason that could explain the central tendency was how some raters found the essays to be repetitive in terms of content and vocabulary, and this made the scoring decision hard. Another reason attributed to the central tendency aspect was how some raters simply needed to be in agreement with their second raters (i.e. the raters they were paired with in scoring)-a point which was well established by Weigle (1994)- as cited in Huang, J. (2009.). In the interviews, two raters said that raters would resort to the middle range of scores to avoid re-reads and, thus, avoid spending more time in finalizing scores.

One final aspect related to scoring behaviors had to do with rater bias-or rather- the characteristic that the raters apparently stressed in their Think Alouds. Kondo-Brown (2002) as cited in Schefer (2008)- said there was some sort of rater bias which different raters exhibited in relation to different criteria. He also said that each rater had a specific pattern of behavior in terms of severity or leniency. In this study though, severity and leniency was only noticed with Raters 3 and 4. Rater 3 gave very low scores along the way. Perhaps this was because she was mostly focused on what the students could not do in her Think Alouds as most of her comments were rather 'negative' reflecting 'frustration' about how students used 'silly' examples or how they failed to apply the rules taught in class. Rater 4, on the other hand, was mostly on the lenient side. Perhaps this was because she would almost always act in empathetic manner and refer to students' effort and give them excuses for mistakes

(i.e. difficult exam or limited time). Barkaoui (2010) mentioned that some raters in his study considered aspects as time limitation when assigning scores. The results of this study, however, not could not be generalized since there were only four raters in the study and only fifteen essays. It would be interesting though to further investigate this finding in terms of how the raters' personal approach and characteristics could impact the scoring behavior and the scores assigned as suggested by Eckes (2008) and Schaefer (2008).

Conclusion

To conclude, instructors exhibited more judgment strategies than interpretation strategies. The rhetorical/ ideational focus was much more attended to compared to language as well. Self-monitoring focus was definitely the least attended to. Perhaps this was because all raters were experienced in this type of scoring. In terms of the scores assigned, the central tendency was mostly exhibited by all raters although one Rater was more generally on the low end of the rubric, whereas another was on the high end of the rubric. The one who was mostly severe was rather negative in her comments in the Think Alouds, whereas the lenient one was mostly sympathetic in her comments while scoring.

Limitations

This study has a number of limitations that researchers need be made aware of. The first limitation is that there were only four participants who took part in the experimental part of scoring fifteen essays while thinking out loud and in the interviews. Those four raters were all experienced, female Egyptian raters working in the same program and department, and this has limited the generalizability of the results of the study on the level of experience, gender, and culture. It would be advisable to use a higher number of raters from different programs, gender, and possibly different nationalities (at least natives and non-natives) to see how different their scoring behavior could be.

Another limitation is that the four raters were requested to work in the scoring and Think Alouds in a difficult time of the semester when

they had their own scoring to do for their own classes and towards the end of the semester. Although they were not rushed to do the scoring or Think Alouds and were clearly told to take the time they needed, the timing might have affected the scoring procedures or the accuracy of assessment.

References

Baker, B.A. (2012). Individual Differences in Rater Decision-Making Style: An Exploratory Mixed-Methods Study. *Language Assessment Quarterly 9* (3), 225-248. https://doi-org.libproxy.aucegypt.edu/10.1080/15434303.2011.637262

Barkaoui, K. (2010). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. TESOL Quarterly 44 (1), 31-57.

_____. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing 28* (1), 51-75. https://doi.org/10.1177/0265532210376379

Cumming, A., Kantor, R., & Powers, D.E. (2001). Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework. (TOEFL Monograph Series No MS-22). Princeton, NJ: Educational Testing Service.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing 25* (2), 155-185. https://doi.org/ 10.1177/0265532207086780

_____. (2012). Operational rater types in writing assessment: Linking Rater Cognition to Rater Behavior. *Language Assessment Quarterly 9*, 270-292. https://doi.org/ 10.1080/15434303.2011.649381

Gebil, A. & Plakans, L. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. *Assessing Writing 21*, 56-73. https://doi-org.libproxy.aucegypt.edu/10.1016/j.asw.2014.03.002

Huang, J. (2009). Factors Affecting the Assessment of ESL Students' Writing. *International Journal of Applied Educational Studies, 5* (1), 1-17

Knoch, U. & Sitajalabhorn, W. (2013). A closer look at integrated writing tasks: Towards a more focused definition for assessment purposes. *Assessing Writing 18* (4), 300-308. https://doi-org.libproxy.aucegypt.edu/10.1016/j.asw.2013.09.003

Lumley, T. (2002). Assessing criteria in a large-scale writing test: What do they really mean to raters? *Language Testing 19* (3), 246-276. https://doi-org.libproxy.aucegypt.edu/10.1191/0265532202lt230oa

Ohta, R. (2018). Integrated listening-to-write assessments: An investigation of score generalizability and raters' decision-making processes (Unpublished doctoral dissertation). University of Iowa, Iowa, USA.

Scahefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing 25* (4), 465-493. https://doi-org.libproxy.aucegypt.edu/10.1177/0265532208094273

Shin, S. Y.& Ewert, D. (2015). What accounts for integrated reading-to-write test scores? *Language Testing 32* (2), 259-281. https://doi-org.libproxy.aucegypt.edu/10.1177/0265532214560257

Wang, D.X. (2014). Beyond CTT and IRT: Using an interactional measurement model to investigate the decision making process of EPT essay raters (Dissertation). University of Illinois at Urbana-Champaign, Illinois, USA. http://hdl.handle.net/2142/49646

Wiseman, C. (2012). Rater Effects: Ego engagement in rater decision-making. *Assessing Writing 17*, 150-173. https://doi-org.libproxy.aucegypt.edu/10.1016/j.asw.2011.12.001

Appendix A (Sample Interview Questions)

- On a continuum of 1-10, 1 most lenient; 10 most severe, where would you place yourself?
- What is your input on the central tendency in grading? Do you do that? Do you think most raters do that? Why/ Why not?
- Do you stick to the analytical rubric or do you treat it holistically? Why/ Why not?
- Do you think that the number of essays we have to read affects our performance? positively or negatively?
- Do you time yourself when you grade an essay? Or a batch?
- Does it make a difference in grading when you know you are placing students to higher levels?
- Do you think the language of rubrics used is vague?
- Do you assign scores in your head as you go along and either confirm or change it along the way? Or do you think about the grade after you are done with the grading?
- What is the sequence of assigning scores? You start with what and end with what?
- Do you follow the order of the rubric or not necessarily?