https://doi.org/10.17323/jle.2022.13466

Experienced and Novice L2 Raters' Cognitive Processes while Rating Integrated and Independent Writing Tasks

Kobra Tavassoli, Leila Bashiri, Natasha Pourdana

Department of ELT, Karaj Branch, Islamic Azad University, Karaj, Iran

ABSTRACT

Background. Recently, there has been a growing interest in the personal attributes of raters which determine the quality of cognitive processes involved in their rating writing practice.

Purpose. Accordingly, this research attempted to explore how the rating experience of L2 raters might affect their rating of integrated and independent writing tasks.

Method. To pursue this aim, 13 experienced and 14 novice Iranian raters were selected through criterion sampling. After attending a training course on rating writing tasks, both groups produced introspective verbal protocols while they were rating integrated and independent writing tasks which were produced by an Iranian EFL learner. The verbal protocols were recorded and transcribed, and their content was analyzed by the researchers.

Results. The six extracted major themes from the content analysis included *content, formal requirement, general linguistic range, language use, mechanics of writing,* and *organization*. The results indicated that the type of writing task (integrated vs. independent) is a determining factor for the number of references experienced and novice raters made to the TOEFL-iBT rating rubric. Further, the raters' rating experience determined the proportions of references they made. Yet, the proportional differences observed between experienced and novice raters in their references were statistically significant only in terms of *language use, mechanics of writing, organization*, and the *total*.

Conclusion. The variations in L2 raters' rating performance on integrated and independent writing tasks emphasize the urgency of professional training to use and interpret the components of various rating writing scales by both experienced and novice raters.

KEYWORDS

cognitive processes, Independent writing task, Integrated writing task, Rating experience, Rubric, Task type, Verbal protocol

Citation: Tavassoli K., Bashiri L., & Pourdana N. (2022). Experienced and Novice L2 Raters' Cognitive Processes while Rating Integrated and Independent Writing Tasks. *Journal of Language and Education, 8*(4), 169-181. https://doi.org/10.17323/jle.2022.13466

Correspondence:
Kobra Tavassoli,
kobra.tavassoli@kiau.ac.ir

Received: December 6, 2021
Accepted: September 15, 2022
Published: December 26, 2022

INTRODUCTION

Rating writing tasks has always been challenging for raters since it often involves subjective evaluation or discriminating judgment (Brown & Abeywickrama, 2010; Leung & Lewkowicz, 2006). This is often the case in scoring second/foreign language (L2) writing tasks that comprise different skills, writing genres, and evaluation criteria (Barkaoui, 2010a; Pourdana et al., 2021). One way to reduce subjectivity in the rating process is to use writing rating scales, which have been the major concern in most large-scale standardized tests. With the growing popularity of these fine-grained rating scales in L2 writing assessment, the L2 researchers' focus has shifted to *how* L2 raters employ such scales and *what* cognitive processes they execute when they rate a piece of writing.

Providing fair and accurate scores to the test-takers' writing is vital for L2 raters because these scores have direct impacts on the future lives of many test-takers who plan to pursue their education at higher levels across the world. Since rating writing tasks is a complex and error-prone process usually performed by human beings of different characteristics (Van Moere, 2014), there is a high demand to study the confluence of raters' attributes in the rating process. One of such important factors which need further investigation is the rater experience. Although some studies have been done to explore the in-depth contribution of the raters' experience in rating L2 writing (Attali, 2016; Barkaoui, 2010a; Davis, 2016; Lim, 2011; Şahan & Razı, 2020; Weigle, 1998), most of them have examined the *actual rating performance* by different raters with marginal attention to the cognitive or psychological processes incorporated into the rating process (Nikmard & Tavassoli, in press). It seems there is a gap in the L2 research literature about the extent to which the nature of cognitive processes can be represented in experienced and novice L2 raters' rating performance on various writing tasks, such as integrated and independent tasks, even though there have been studies on test takers' differing performance on these task types (e.g., Ahmadi & Mansoordehghan, 2015; Plakans, 2010; Shi et al. 2020). This study, therefore, employed the verbal protocol method of introspection to collect data on the underlying cognitive processes of L2 raters with varying degrees of experience (experienced vs. novice) while they engage in rating the test-takers' performance on integrated and independent writing tasks. The results of this study can have important implications for L2 raters and rater trainers to raise their awareness of the game-changing cognitive processes enacted in their rating performance.

Task-Based Assessment of L2 Writing

The quality of writing in L2 brings about enormous advantages to students such as showing their academic character, promoting effective communication, advancing their higher-order thinking skills, making logical and convincing arguments, demonstrating their ideas and re-assessing them, and promoting them to their future careers (Beck et al., 2018; Swales & Feak, 2004). Writing is probably the most complex L2 skill to teach and to assess in most EFL contexts such as Iran, unless the L2 teacher is experienced enough to manage the dilemma (Hyland, 2003; Klimova, 2013). Any teaching practice in a formal setting is accompanied by sequential and/or subsequent assessment. Accordingly, the teaching and assessment of writing are not exceptional. Writing can be assessed through various writing tasks which are vastly different in terms of their focus, the type of challenge they generate, the feedback type they provide to L2 writers, and their degree of correspondence to real-world tasks (i.e., authenticity) that L2 learners wish to perform. The outcome of assessing L2 writing is usually a gain score, which is the by-product of the dynamic interactions among the writer, the writing task, the written product, the rater, and the rating procedure.

Traditionally, the dominant writing task in large-scale international tests such as the International English Language Testing System (IELTS) or Test of English as a Foreign Language (TOEFL) has been the independent writing task. Defined by the TOEFL Family of Assessments (https://www.ets.org/toefl), an *independent* writing task requires L2 learners to draw on their personal experience, opinion, and knowledge when responding to a prompt. They are called independent tasks because the L2 learner alone is the source of information. However, independent writing tasks are argued for being decontextualized as students may not have any information about the topic to write about; also these tasks do not let students benefit from other available resources (Ahmadi & Mansoordehghan, 2015). That is why, more recently, integrated writing tasks have been accommodated to international tests.

The TOEFL Family of Assessments (https://www.ets.org/toefl) defines an *integrated* writing task as a task that demands test takers to integrate the database from various resources when responding to a prompt. In these tasks, test takers are required to read a short passage and/or listen to an academic lecture and write the response to the prompts by using the information incorporated into the passage and/or the talk. It is a common belief that integrated writing tasks are more contextualized and authentic (Ahmadi & Mansoordehghan, 2015). However, except few studies such as Michel et al. (2020) and Uludag et al. (2021), not much research has been done to cross-examine the L2 learners' performance on independent and integrated writing tasks or to analyze L2 raters' cognitive processes when they rate either of these tasks. Since performance on these two tasks requires different cognitive engagement (relying on one's own knowledge in independent tasks versus incorporating information from other sources in integrated tasks), it was presumed that L2 raters' cognitive processes may also differ when rating these two types of tasks. Therefore, these two popular task types were selected for further investigation in this study.

Scoring Rubrics and Rating Scales

Brookhart (2013) defined a rubric as a logical and clear-cut set of criteria to evaluate students' language production. A scoring rubric also includes precise descriptions of the performance levels that match those criteria. Relying on standard scoring rubrics, L2 teachers can provide informative feedback to their students by locating the problems in their output, identifying their errors, and providing diagnostic information about their strengths and weaknesses (Suskie, 2008). Various scoring rubrics are usually well-tuned to specific purposes. The most important purpose of selecting a scoring rubric is to evaluate performance, either while the student is producing language or after the language product or the task outcome is ended. Rubrics can also shape the teacher's/rater's rating behaviors. In a classroom assessment situation, for example, objective judgments can take place by corresponding the teacher/rater observation of a student's work to the descriptions embedded in the scoring rubric. The quality judgment based on a standard scoring rubric can subsequently be employed in terms of diagnostic feedback or formative assessment by L2 teachers and/or raters (Brookhart, 2013).

Rating scales are divided into three major types of *primary trait, holistic*, and *analytic* (Suskie, 2008). Suskie's classification is founded upon the binary aspects of rating scales; whether they are specific to a single task or several tasks and whether a single score or several scores are granted to each writing product. The primary trait is well-recognized by most task developers which enable L2 raters, teachers, and students to concentrate on a single characteristic of the task, such as appropriate text staging, creative response, reference to sources, and so on (Weigle, 2002). The primary trait rubric is usually used in assessing the test-takers' basic writing skills. The holistic rating scale has been widely used in various writing assessment programs and international tests over the past 25 years. It serves test-takers with a single unified score that summarizes the scoring criteria. The goal of this scale is to evaluate a writer's total proficiency through the quality of a given writing sample (Hyland, 2003). Finally, the analytic rating scale evaluates a piece of writing based on the microscopic features or linguistic criteria such as vocabulary, grammar, content, organization, cohesion, or mechanics of writing.

In this study, aligned to several other studies (Attali, 2016; Hyland, 2003; James, 2006; Shi et al., 2020; Zanders & Wilson, 2019), we adopted the holistic rating scale to assess independent and integrated writing task outcomes. According to Harsch and Martin (2013), one of the major pros of holistic rating scales is that they concentrate on the strong points of a written product rather than its flaws and weaknesses. Yet, the single rounded scores that holistic rating scales compose eradicate the chances of L2 teachers or raters to discriminate certain lower-level skills of writing such as rhetorical features, choice of words, or mechanics of writing. Neither do they offer substantial diagnostic information on L2 learners' task outcomes. Overall, using holistic rating scales is more time-saving and manageable than primary or analytic rating scales to most L2 teachers and raters.

Process of Rating Writing Tasks

As a critical step in the assessment process, rating writing connects the test-takers' writing performance to the descriptors in the rating scale. In other words, in the rating process, the attributes of a written product are converted into a rating that measures the extent to which the scale descriptors have been realized. Various factors can determine the reliability of a rating, including the raters' linguistic background, professional background, cognitive processes, gender, and rating experience. Rating experience is one of the most important *rater effects* which directly impact rating writing tasks (Davis, 2016; Duijm et al., 2018; Lim, 2011). Furthermore, rater training can also impact the process of rating writing. In training sessions, instructions on various rating scales are usually provided so that raters can perform the rating process systematically and consistently. A typical rating training session can be handled face-to-face or online through workshops or webinars (Attali, 2016). In rating training sessions, novice raters review the writing prompts, scoring rubrics, rating scales, and the benchmark written responses, and consult over controversial issues with more experienced raters. Their training is evaluated by rating sample responses and receiving feedback on their assigned scores from experienced raters. Finally, prospective L2 raters should pass a certification test to receive authorization to rate writing tasks. The process of rating writing tasks may also be affected by other factors such as the type of task or the scoring method. Recently, Khodi (2021) in a G-theory analysis of rater, task, and scoring method examined the affectability of writing assessment scores. Using various raters, tasks, and scoring methods, he found that to reach maximum generalizability, students should take two writing tasks and their performance should be evaluated by at least four raters using at least two scoring methods. In other words, a single rating of a single performance by a single rater cannot be trusted because of the subjectivity involved in the process of rating writing tasks.

The L2 research, however, has documented little evidence on the usefulness of rating training programs and the certification procedures or the potential impact of rating experience to determine the cognitive processes raters are involved in while rating writing tasks. Some researchers supported the positive impact of the rating training on lowering the rater subjectivity in terms of severity or leniency and enhancing their consistency in scoring (Elder et al., 2007; Fahim & Bijani, 2011; Weigle, 1998). On the other hand, several studies speculated the constructive role of training in eliminating rater variability by evidence of the recorded variance in experienced raters' assigned scores to a certain written performance (Eckes, 2012; Long & Pang, 2015). More importantly, there is a scarcity of research on L2 raters' cognitive processes they are involved in while rating different writing tasks (Nikmard & Tavassoli, in press). Barkaoui (2010a), for instance, examined the role of the rating scale and rating experience, and the variability they would cause in the rating process of an L2 essay. The verbal protocol method of introspection was carried out to investigate the roles of the rating experience, rating scales, and their interactions on raters' decision-making processes. He found that the type of rating scale had larger effects than the rating experience on the raters' rating processes. In another study, Barkaoui (2010b) cross-examined experienced and novice raters in their holistic and analytic scoring performance. The results showed that both groups prioritized the communicative quality of the writings. Yet, experienced raters were more severe to the linguistic accuracy than novice raters who were more critical to the argumentative voice of the writers.

To void the gap in the L2 literature on rating writing tasks, therefore, this study adopted a cognitive approach to the study of rater variability and aimed to analyze the differences between experienced and novice L2 raters in terms of the cognitive processes they incorporate into rating integrated and independent writing tasks. In this regard, the

TOEFL-iBT scoring rubric was used by both experienced and novice raters to rate integrated and independent writing tasks. Even though both experienced and novice raters in this study were familiar with rating mock TOEFL-iBT writing tasks, to ensure consistency in their rating, they attended a rating training session to get more information about the process of rating writing tasks in general and rating TOEFL-iBT integrated and independent writing tasks in particular. More details are provided in the procedure section. Accordingly, to serve the objectives of this study, two research questions were raised: (1) What difference does the type of writing task (integrated vs. independent) make on the rating performance of experienced and novice L2 raters? (2) What difference does the rating experience (experienced vs. novice) make on the L2 raters' rating of integrated and independent writing tasks?

METHODS

Participants

Since the focus of this study was on L2 raters' cognitive processes in rating writing tasks, the participants were 27 Iranian raters who were L2 speakers of English. They were selected through criterion sampling where only those who meet the researchers' predetermined criteria are selected (Dörnyei, 2007). This sampling was used since being an L2 rater of writing tasks was a prerequisite for the completion of this study. The researchers invited the participants who met this criterion from different language institutes to take part in this study. The selected participants were 13 experienced raters (six females and seven males) and 14 novice raters (10 females and four males) whose educational background was Master or Ph.D. in teaching English as a foreign language (TEFL), English literature, or English translation. All the raters agreed willingly to participate in this study. Initially, 30 raters (15 experienced and 15 novices) had agreed to participate in the study, however, when the study began, three raters (2 experienced and 1 novice) withdrew from the study. To check the suitability of the sample size, a prior power analysis was conducted (Hoenig & Heisey, 2012). Accordingly, the sample size of 15 in this study was acceptable to retain the 80% power at p = .05.

The trait of *rating experience* was operationally defined as having over five years of teaching EFL and rating experience for the 13 experienced raters. On the other hand, the 14 novice raters were those who had less than three years of teaching EFL and rating experience.

There was also a randomly chosen Iranian EFL learner who agreed willingly to participate in this study. She completed one integrated and one independent writing task from a mock TOEFL-iBT test to be scored by the 27 raters. The informant was a 25-year-old female undergraduate student who had been studying English for seven years in a language institute at the time of this research. The informant's mean score (M) on the integrated writing task was 3.55 (M of experienced raters = 3.46; M of novice raters = 3.64), and her mean score (M) on the independent writing task was 3.14 (M of experienced raters = 3; M of novice raters = 3.28). Overall, the informant's writing mean score from all the raters' scores was 3.34. When converted based on TOEFL-iBT score conversion tables (Gallagher, 2005), the informant's writing score changed to 22. The rationale for selecting only one informant was to ensure the rater participants would provide rich and detailed introspection while producing verbal protocols on their rating performance. However, the researchers acknowledge that having only one sample for each integrated and independent writing task would not be representative enough and would jeopardize the generalizability of the findings. Nevertheless, the EFL learner was chosen randomly to alleviate this problem as much as possible.

Instruments

Integrated and Independent Writing Tasks

The informant was asked to write one integrated and one independent writing essay prompted in a mock TOEFL-iBT test taken from Gallagher (2005). Prompt 1 was an integrated task that required the informant to read a passage, listen to a lecture about the earthworms and other soil dwellers, and describe *the problems caused by earthworms in the forest ecosystems* by explaining how these problems contradicted the information in the reading. The allotted time was 20 minutes for drafting an essay of around 150–225 words. Prompt 2 was an independent writing task that required the informant to write an expository essay on *the importance of what we learn inside the school and what we learn outside the school*, based on her knowledge and experience. The allotted time was 30 minutes for drafting an argumentation with a minimum of 300 words.

TOEFL-iBT Scoring Rubric

The TOEFL-iBT writing rubric, which was used in this study, consists of four components of language use (i.e., how well the examinee can use grammar and vocabulary), organization (i.e., how well the examinee can put the sentences into a logical order), clarity (i.e., how clear, concise, and ready to be read the examinee's writing is), and development (i.e., how coherent the examinee's essay is) on a 6-band scale (ranging from 0 to 5) (https://www.ets.org/s/toefl/pdf/toefl_writing_rubrics.pdf).

Introspective Verbal Protocol

In this study, the verbal (think-aloud) protocol was used as the method of data collection. As a methodological tool, the verbal protocol is a model of information processing based on the verbalization of inner speech. Introduced by Ericsson and Simon (1993), the verbal protocol is a common technique to ask individuals to vocalize introspectively what is going through their minds as they are solving a problem or performing a task. Verbal reporting allows researchers to explore how individuals can be different in their approach to a certain problem (Krahmer & Ummelen, 2004). This technique was used since it is one of the most common ways of exploring the participants’ mental or cognitive processes when they perform a certain task (Dörnyei, 2007).

Procedure

After the sample of 27 experienced and novice Iranian raters were selected, the purpose and the procedure of the study were explained to them. Then, they attended a two-session training course on rating writing tasks tutored by an experienced rater who had 14 years of experience in rating mock TOEFL-iBT writing tests. In the first 90-minute session, an introduction was made to writing tasks, the rating process, the TOEFL-iBT writing rubric and its components, and the procedure of the verbal protocol. In the second 90-minute session, the participants practiced rating four independent and four integrated writing tasks which were selected from TOEFL-iBT writing sample responses. The participants justified their assigned scores to the tasks in a post-rating discussion which was followed by comparing their scores to the designated scores by the TOEFL-iBT examiners.

Shortly after the rating tutorial, the informant completed two writing tasks of a mock TOEFL-iBT test which lasted for 50 minutes. Her task outcomes were distributed to the experienced and novice raters. They were required to use the 6-point TOEFL-iBT rating scale by assigning a holistic score to each writing task while they were introspectively producing verbal protocols on their rating process. The verbal protocols were recorded, transcribed, and inserted into QSR NVivo version 10. The recorded verbal protocols while rating the independent and integrated tasks were then subjected to content analysis. The process of content analysis was carried out by the researchers collaboratively to reach a full consensus.

Coding System

The contents of the raters’ verbal protocols on rating the independent and integrated writing tasks were pooled and encoded to extract the most frequent themes and subthemes representing the criteria in TOEFL-iBT writing rating rubrics. The researchers collaboratively developed a coding system with six major themes, including *Content, Formal requirement, General linguistic range, Language use, Mechanics of writing,* and *Organization*, following TOEFL-iBT writing rating rubrics. Each of these major themes also consisted of several subthemes for each task. The subthemes were basically extracted from the raters’ verbal protocols while rating each task. Table 1 presents the six major themes along with their subthemes and sample examples from the raters’ verbal protocols. As it can be seen in Table 1, many of the subthemes of the integrated and independent writing tasks were similar. However, there were some differences in the subthemes of the two writing tasks which are notified in Table 1.

RESULTS

Proportional Distribution of Themes/ Subthemes in the Writing Tasks

Tables 2 and 3 represent the encoded themes, the respective subthemes, and their proportional distribution for the experienced and novice raters on the integrated and independent writing tasks, respectively. Illustrated in Table 2, on rating the integrated writing task, experienced raters made more references to the scoring rubric than novice raters (165 to 115, respectively). The experienced and novice raters’ focus was mutually on the theme of *Language use: Structure* ($f_{\text{experienced}} = 40$, $f_{\text{novice}} = 20$), before their attention was shifted to the theme of *Content: Making connections between the passage and the lecture* ($f_{\text{experienced}} = 17$, $f_{\text{novice}} = 11$) and *Content: Selecting the important information from the lecture* ($f_{\text{experienced}} = 13$, $f_{\text{novice}} = 14$). As a point of departure, novice raters paid more attention to the subthemes of *General linguistic range: Accuracy, Clearness, Preciseness* (f = 17) than experienced raters who focused more intensively on the themes of *Mechanics of writing* (f = 7) and *Organization*, with special attention to its subtheme of *Using a concluding paragraph* (f = 10).

As Table 3 displays, on rating the independent writing task, both experienced and novice raters made more references to the rubric. Moreover, experienced raters had a higher record than novice raters (208 to 156, respectively). The experienced and novice raters focused mostly on the theme of *Language use: Structure* ($f_{\text{experienced}} = 45$, $f_{\text{novice}} = 28$) and *Language use: Vocabulary* ($f_{\text{experienced}} = 26$, $f_{\text{novice}} = 15$). They also showed rather similar interests in the theme of *Organization* ($f_{\text{experienced}} = 72$, $f_{\text{novice}} = 45$). Further, the theme of *General linguistic range: Accuracy* ($f_{\text{experienced}} = 0$, $f_{\text{novice}} = 2$)*, Clearness* ($f_{\text{experienced}} = 4$, $f_{\text{novice}} = 5$)*,* and *Preciseness* ($f_{\text{experienced}} = 0$, $f_{\text{novice}} = 1$) was the least noticed theme by both experienced and novice raters. Similar to rating the integrated writing task, experienced raters showed more interest than novice raters to *Mechanics of writing* ($f_{\text{experienced}} = 11$, $f_{\text{novice}} = 1$).

Analysis of Verbal Protocols on Rating the Writing Tasks

The total records of themes extracted from the verbal protocols on rating the integrated writing task were 280, of which 165 records were made by experienced raters and 115 by novice raters. Table 4 displays the proportional theme dis-

Table 1

Themes, Subthemes, and Examples from the Raters' Verbal Protocols

Theme	Subtheme	Example
Content	Addressing all aspects of the topic	In terms of task achievement, I can say that the examinee has done a good job.
	Conveying the message	The writer was successful to convey the message.
	Relevance to the prompt	The writing is not directly dealing with the question raised in the topic.
	Comprehending the passage and the lecture*	The comprehension regarding the listening or the lecture was quite well.
	Making connections between the passage and the lecture*	There isn't a clear connection between the points she made and the points made in the passage and the lecture.
	Expressing the main idea*	She was successful in expressing the main ideas.
	Selecting the important information from the lecture*	Some important parts from the lecture or the reading passage have been selected.
	Convincing the reader**	The reasons are not convincing or do not persuade the reader.
	Using exemplification**	Examples are not developed well.
	Using explanation**	Not sufficient explanations or details are provided.
Formal requirement	Number of paragraphs	She wrote just two paragraphs.
	Number of words	The number of words is really few here.
General linguistic range	Accuracy	There is some inaccuracy.
	Clearness	Some sentences are vague.
	Preciseness	Everything is precise.
Language use	Structure	Grammatical errors are noticeably present.
	Vocabulary	She had a good command of vocabulary.
Mechanics of writing	Punctuation	There is a lack of punctuation.
	Spelling	There is no important misspelling.
	Capitalization*	There are problems with capital letters.
Organization	Logical order of ideas	The writer should have first mentioned the problematic areas.
	Using a topic sentence	There is the absence of a well-developed thesis statement.
	Using supporting sentences	There is a lack of supporting ideas.
	Coherence	No proper connection is seen between the sentences.
	Good organization	The writer is clearly not familiar with the way to organize a piece of writing.
	Using an introductory paragraph	The introduction is missing.
	Using body paragraphs	Everything is right about the body paragraphs.
	Using a concluding paragraph	The concluding part could be more academically written.
	Development of ideas**	The development of ideas is good.
	Development of paragraphs**	There is limited development of the paragraphs.

*Note. * Specific to the integrated writing task*

*** Specific to the independent writing task*

Table 2

Proportional Distribution of Themes/Subthemes in Rating the Integrated Writing Task (f=frequency)

Theme	Subtheme	Experienced Raters' f	Novice Raters' f
Content	Addressing all aspects of the topic	2	0
	Conveying the message	5	8
	Relevance to the prompt	3	6
	Comprehending the passage and the lecture	2	0
	Making connections between the passage and the lecture	17	11
	Expressing the main idea	2	3
	Selecting the important information from the lecture	13	14
Formal requirement	Number of paragraphs	4	2
	Number of words	6	4
General linguistic range	Accuracy	2	7
	Clearness	8	7
	Preciseness	1	3
Language use	Structure	40	20
	Vocabulary	10	6
Mechanics of writing	Punctuation	1	1
	Spelling	5	0
	Capitalization	1	0
Organization	Logical order of ideas	5	2
	Using a topic sentence	5	1
	Using supporting sentences	2	1
	Coherence	8	10
	Good organization	5	5
	Using an introductory paragraph	7	0
	Using body paragraphs	1	0
	Using a concluding paragraph	10	4
Total		**165**	**115**

tribution for the experienced and novice raters in rating the integrated writing task along with the related chi-square values comparing the frequencies of each theme and the total. The level of significance was set at α = .05.

Displayed in Table 4, in the theme of *Content*, there was a small difference between the records of the two rater groups with insignificant chi-square measure (χ^2 = .04, p = .82 > .05). The themes of *Formal requirement* (χ^2 = 1.00, p = .31 > .05) and *General linguistic range* (χ^2 = 1.28, p = .25 > .05) likewise were recorded with small and insignificant differences by experienced and novice raters. On the other hand, experienced and novice raters had considerable differences in recording the themes of *Language use* (χ^2 = 7.57, p = .00 < .05), *Mechanics of writing* (χ^2 = 4.5, p = .03 < .05) and *Organization* (χ^2 = 6.06, p = .01 < .05), with experienced raters having almost twice more records than novice raters. Further, there was a significant difference between experienced and novice raters' records regarding the total themes they mentioned when rating the integrated writing task (χ^2 = 8.93, p = .00 < .05).

The other analysis was carried out on the content of verbal protocols produced by experienced and novice raters while rating the independent writing task. Here, the accounts of themes were 364, of which 208 belonged to experienced raters and 156 to novice raters. In Table 5, the proportional distribution of the themes produced by experienced and novice raters for the independent task was compared by running another set of chi-square tests.

Table 3

Proportional Distribution of Themes/Subthemes in Rating the Independent Writing Task (f=frequency)

Theme	Subtheme	Experienced Raters' f	Novice Raters' f
Content	Addressing all aspects of the topic and the task	10	12
	Conveying the message	1	4
	Relevance to the prompt	1	1
	Convincing the reader	1	2
	Using exemplification	11	12
	Using explanation	12	12
Formal requirement	Number of paragraphs	5	3
	Number of words	9	14
General linguistic range	Accuracy	0	2
	Clearness	4	5
	Preciseness	0	1
Language use	Structure	45	28
	Vocabulary	26	15
Mechanics of writing	Punctuation	7	1
	Spelling	4	0
Organization	Logical order of ideas	1	2
	Using a topic sentence	7	3
	Using supporting sentences	5	1
	Coherence	15	14
	Good Organization	11	10
	Using an introductory paragraph	5	1
	Using body paragraphs	2	0
	Using a concluding paragraph	9	3
	Development of ideas	10	10
	Development of paragraphs	7	0
Total		**208**	**156**

In Table 5, a similar pattern of distribution can be seen between the records of themes extracted from experienced and novice raters' verbal protocols. Accordingly, small and statistically insignificant differences were observed between experienced and novice raters on the themes of *Content* ($\chi^2 = .62$, $p = .43 > .05$), *Formal requirement* ($\chi^2 = .29$, $p = .59 > .05$), and *General linguistic range* ($\chi^2 = 1.33$, $p = .24 > .05$). However, experienced raters had a meaningful difference in their much higher records of references than novice raters to the theme of *Language use* ($\chi^2 = 6.87$, $p = .00 < .05$), *Mechanics of writing* ($\chi^2 = 8.33$, $p = .00 < .05$) and *Organization* ($\chi^2 = 6.75$, $p = .00 < .05$). Once again, there was also a significant difference between the total records of themes experienced and novice raters provided when rating the independent writing task ($\chi^2 = 7.43$, $p = .00 < .05$).

Next, to check whether rating experience is associated with the type of writing task raters rate, a Correspondence Analysis was run. This is a multivariate technique to discover the relationships among categorical variables in graphical form (Zabihi et al., 2019). After identifying a link between the levels of the two categorical variables (rating experience and writing task type), a Correspondence Analysis was run by determining 2 dimensions corresponding to the two variables. Figure 1 shows the result of this analysis. Dimension 1 refers to the rating experience (experienced vs. novice) and Dimension 2 refers to the writing task type (integrated vs. independent). As it can be seen in Figure 1, there is a clear distinction between the two groups of raters (novice vs. experienced). This is a reconfirmation of the results of Tables 4 and 5 which showed a significant difference between the

Table 4

Theme Distribution in Rating the Integrated Writing Task by Experienced and Novice Raters (f=frequency)

Theme	Experienced Raters' f	Novice Raters' f	χ^2	p
Content	44	42	.04	.82
Formal requirement	10	6	1.00	.31
General linguistic range	11	17	1.28	.25
Language use	50	26	7.57	.00*
Mechanics of writing	7	1	4.5	.03*
Organization	43	23	6.06	.01*
Total	**165**	**115**	**8.93**	**.00***

Table 5

Theme Distribution in Rating the Independent Writing Task by Experienced and Novice Raters (f=frequency)

Theme	Experienced Raters' f	Novice Raters' f	χ^2	p
Content	36	43	.62	.43
Formal requirement	14	17	.29	.59
General linguistic range	4	8	1.33	.24
Language use	71	43	6.87	.00*
Mechanics of writing	11	1	8.33	.00*
Organization	72	44	6.75	.00*
Total	**208**	**156**	**7.43**	**.00***

two rater groups' rating of both writing tasks. However, rating both integrated and independent writing tasks is associated with rating experience. In other words, both novice and experienced raters rated either of the two writing tasks similarly.

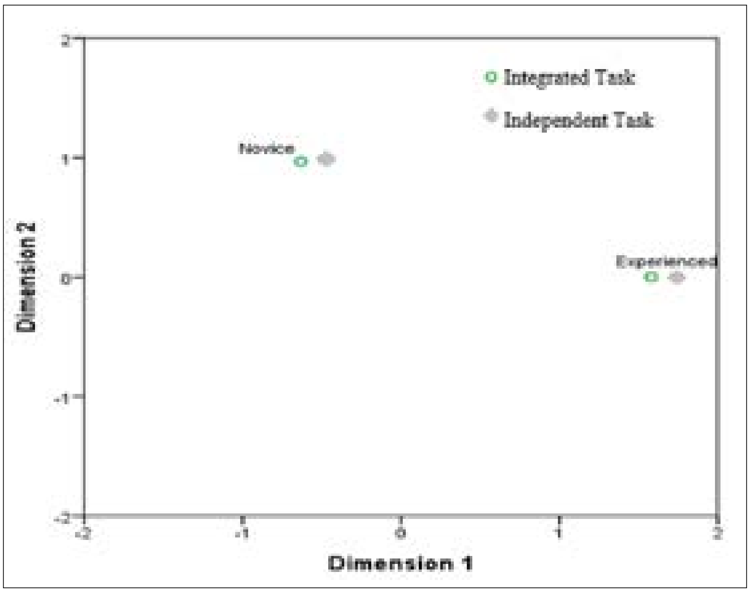
DISCUSSION

The purpose of this research was to explore the potential interplay between the rating experience and the type of writing task to determine the nature of L2 raters' cognitive processes involved in rating integrated and independent writing tasks. The results of the first research question in this study showed that the type of writing task (integrated vs. independent) was a determining factor in the number of references experienced and novice raters made to the TOEFL-iBT writing rating rubric. In other words, both experienced and novice raters produced higher records of references to the TOEFL-iBT rating rubric on rating the independent writing task than the integrated writing task, where experienced raters' records outnumbered novice raters' records. Moreover, the results of the second research question indicated that the rating experience was another determining factor in the proportion of references L2 raters made on both types of tasks. In other words, while experienced and novice raters incorporated similar cognitive processes (represented by the six major extracted themes) in rating the integrated and independent writing tasks, they had meaningful differences by the larger proportion of references experienced raters made to the themes of *Language use, Mechanics of writing,* and *Organization*, as well as the total themes on both tasks.

The first research question is discussed in terms of the heavy cognitive load independent writing tasks might cause which makes them more challenging for L2 learners to write about and subsequently for L2 raters to rate. This cognitive load to construct the textual meaning has been labeled as task representation by Wolfersberger (2007). In other words, when students face a writing task, they need to understand which skills, products, and processes the task requires and they should plan to prepare a written product that properly matches the task (Wolfersberger, 2007). Several studies suggested that task representation becomes much easier to fulfill when the L2 writer/learner has access to external resources as the writing task input that is the case in integrated tasks (Allen, 2004; Plakans, 2010; Ruiz-Funes, 2001; Wolfersberger, 2007), where the main concerns for the L2 writer/learner are how to employ resource texts in their writing and how to reiterate them appropriately (Plakans, 2010). Hence, it seems that integrated writing tasks become less challenging to L2 learners and raters. This was confirmed in this study as L2 raters provided less records in their ver-

Figure 1

Joint distribution of rating experience and writing task type



Dimension 1: Rating experience (experienced vs. novice);

Dimension 2: Writing task type (integrated vs. independent)

bal protocols while rating the integrated writing task which might mean that rating the integrated task requires fewer cognitive demands on raters as it is more content-controlled and less creative.

The findings on the first research question corroborated some previous studies (Allen, 2004; Wolfersberger, 2007) but partially contradicted several others (Ahmadi & Mansoordehghan, 2015; Michel et al., 2020; Plakans, 2010; Uludag et al., 2021). The results of the first research question did not match those of Ahmadi and Mansoordehghan (2015) since in contrast to this study, they found that task type (independent vs. integrated) did not have a significant effect on the students' writing performance. However, if test takers' cognitive processes rather than their writing performance are investigated while they complete independent and integrated tasks, differences may be found as it was the case regarding the raters' cognitive processes while they rated these two types of writing tasks in this study. Moreover, the results of this study were in contrast to Michel et al. (2020) and Uludag et al. (2021) who found that integrated writing tasks are more challenging for test takers and elicit more dynamic and varied behaviors and cognitive processes in test takers. This contradiction might be due to the nature of the participants in Michel et al.'s and Uludag et al.'s studies who were test takers versus the participants in this study who were raters. Most probably, raters and test takers go through different cognitive processes while dealing with writing tasks. Although writing integrated tasks may be more demanding for test takers since they have to integrate information from different sources, it seems that rating independent tasks requires more energy and mental processing in L2 raters. Also, dealing directly with the participants' cognitive processes through verbal protocols may be another important issue which resulted in differences between the findings of this study and other studies since verbal protocols chiefly explore mental processes, rather than speculating about them.

The second research question is discussed by the argument of the rater subjectivity and the potential interaction between rating experience, rater training, and rater severity/leniency. The term rater severity refers both to the general tendency of a rater to assign higher or lower ratings than the average raters, and to the observed differences among raters in terms of their interpretations of the rating rubrics (Lim, 2011). Rater severity is an integral component of rater biasedness which might be nurtured by rater experience or rater training (Eckes, 2011; Myford & Wolfe, 2003). However, as Eckes (2011) emphasized, there is a lack of research on factors affecting rater severity. Despite this lack of research, Khodi (2021) investigated the issue of rater experience and rater severity and their impact on writing scores and suggested that the test takers' writing performance should be rated by at least four raters using at least two scoring methods to avoid rater biasedness. Similarly, the findings of this study also recommend that since rater experience is an influential factor in the cognitive processes raters engage in when rating both integrated and independent writing tasks, asking raters with different levels of experience to rate writing tasks is necessary to have a better evaluation of the test takers' writing performance and to avoid rater severity and biased scoring. Moreover, paying attention to what cognitive processes experienced L2 raters focus on while rating writing tasks and inform novice L2 raters about such processes in training sessions can be quite helpful. Yet, further studies are demanded on this venue of research.

Furthermore, although some research show that rating experience influences raters' scoring performance both in terms of leniency and focus (Duijm et al., 2018), some other research show that the type of rating rubric (analytic vs. holistic) (Barkaoui, 2010a), the depth of learning that happens in the process of training (Attali, 2016), and the text quality (Şahan & Razı (2020) are more determining factors than rating experience in the raters' unbiased rating and decision-making processes. In a nutshell, factors such as the rating rubric, rater training, rater knowledge, and text quality are as influential as if not more influential than rating experience in the scores raters assign to writing tasks. Therefore, following the results of different studies as well as this study, both rater experience and training should be regarded as important factors to consider when studying raters' scoring of writing tasks.

CONCLUSION

This study revealed that for L2 raters, their rating practice would be affected by the interplay of their rating experience and the type of writing task. Therefore, the findings of this study have several pedagogical implications. To reduce the rater variability and bias in the rating process, the most common solution is rater training, where L2 raters with various levels of rating experience are (re)introduced to the rating criteria followed by their immediate and delayed rating practice to safeguard the sustainability of the training. Further, since the standard rubrics most likely make the rating more reliable, raise clarity in rater judgment, and lessen rater subjectivity, L2 raters and teachers should receive the how-to instructions on using rubrics both as a grading and teaching device. Moreover, by employing verbal protocols or stimulated recalls as a pedagogical tool, L2 raters and teachers might raise in students the kind of self-awareness they need to engage in their writing process. Finally, making L2 learners familiar with instructions to rating rubrics can help them improve not only their self-directness but also their writing ability.

The findings of this study should be recognized in light of some limitations. One major limitation was the sample size. In this study, a single EFL learner's writing performance on integrated and independent writing tasks was rated by 27 raters in a one-shot comparative research. This research can be replicated by rating more writing samples from different EFL learners in an extended period of time to enhance the generalizability and sustainability of the findings. Another limitation of the study was not considering the text length in integrated and independent writing tasks. However, text length might be an influential factor in the number of records both experienced and novice raters provided. In future studies, this factor should also be investigated. Moreover, since various rater differences such as their educational background can infiltrate the findings of the study, their inclusion is highly recommended in future research. Also, the use of introspective verbal protocols has certain methodological limitations. It is a complex technique that may affect the raters' performance by causing distractions, stress, and low task representation, which eventually affect the transferring of results to natural rating contexts. Therefore, to remedy the shortcomings of using verbal protocols, it can be empowered with other techniques, such as interviews or stimulated recalls, which adopt a more emic approach to data collection by retrieving the raters' self-evaluation of their rating performance.

DECLARATION OF COMPETING INTEREST

None declared.

AUTHOR CONTRIBUTION STATEMENT

K.Tavassoli: Conceptualization, Methodology, Supervision, Visualization, Writing- Original draft preparation, Writing- reviewing and editing.

L. Bashiri: Conceptualization, Data curation, Formal analysis, Investigation, Project administration, Software.

N. Pourdana: Conceptualization, Methodology, Visualization, Writing- Original draft preparation, Writing- reviewing and editing.

REFERENCES

Ahmadi, A., & Mansoordehghan, S. (2015). Task type and prompt effect on test performance: A focus on IELTS academic writing tasks. *Journal of Teaching Language Skills, 6*(3), 1-20. http://doi.org/10.22099/jtls.2015.2897.

Allen, S. (2004). Task representation of a Japanese L2 writing and its impact on the usage of source text information. *Journal of Asian Pacific Communication, 14*(1), 77-89. http://doi.org/10.1075/japc.14.1.06all.

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing, 33*(1), 99-115. https://doi.org/10.1177/0265532215582283.

Barkaoui, K. (2010a). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*(1), 54-74. https://doi.org/10.1080/15434300903464418.

Barkaoui, K. (2010b). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly, 44*(1), 31-57. http://doi.org/10.2307/27785069.

Beck, S. W., Llosa, L., Black, K., & Anderson, A. T. (2018). From assessing to teaching writing: What teachers prioritize. *Assessing Writing, 37*, 68-77. https://doi.org/10.1016/j.asw.2018.03.003.

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*: Alexandria.

Brown, H. D., & Abeywickrama, P. (2018). *Language assessment: Principles and classroom practices* (3rd ed.). Pearson Education.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117-135. https://doi.org/10.1177/0265532215582282.

Dörnyei, Z. (2007). *Research methods in applied linguistics*. Oxford University Press.

Duijm, K., Schoonen, R., & Hulstijn, J. H. (2018). Professional and non-professional raters' responsiveness to fluency and accuracy in L2 speech: An experimental approach. *Language Testing, 35*(4), 501-527. https://doi.org/10.1177/0265532217712553.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments.* Peter Lang.

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly, 9*(3), 270-292. https://doi.org/10.1080/15434303.2011.649381.

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. https://doi.org/10.1177/0265532207071511.

Ericsson, K. A., & Simon, H. (1993). *Protocol analysis: Verbal reports as data.* MIT Press.

Fahim, M., & Bijani, H. (2011). The effects of rater training on raters' severity and bias in second language writing assessment. *International Journal of Language Testing, 1*(1), 1-16.

Gallagher, N. (2005). *Delta's key to the next generation TOEFL test: Advanced skill practice for the IBT*. Delta Publishing Company.

Harsch, C., & Martin, G. (2013). Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice, 20*(3), 281-307. https://doi.org/10.1080/0969594X.2012.742422.

Hoenig, J. M., & Heisey, D. M. (2012). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistican, 55*(1), 19-24. https://doi.org/10.1198/000313001300339897.

Hyland, K. (2003). *Second language writing*: Cambridge University Press.

James, C. L. (2006). Validating a computerized scoring system for assessing writing and placing students in composition courses. *Assessing Writing, 11*(3), 167-178. https://doi.org/10.1016/j.asw.2007.01.002.

Khodi, A. (2021). The affectability of writing assessment scores: A G-theory analysis of rater, task, and scoring method contribution. *Language Testing in Asia*, *11*(30), 1-27. https://doi.org/10.1186/s40468-021-00134-5.

Klimova, B. F. (2013). Developing thinking skills in the course of academic writing. *Procedia-Social and Behavioral Sciences, 93*, 508-511. https://doi.org/10.1016/j.sbspro.2013.09.229.

Krahmer, E., & Ummelen, N. (2004). Thinking about thinking aloud: A comparison of two verbal protocols for usability testing. *IEEE Transactions on Professional Communication, 47*(2), 105-117. https://doi.org/10.1109/TPC.2004.828205.

Leung, C., & Lewkowicz, J. (2006). Expanding horizons and unresolved conundrums: Language testing and assessment. *TESOL Quarterly, 40*(1), 211-234. https://doi.org/10.2307/40264517.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing, 28*(4), 543-560. https://doi.org/10.1177/0265532211406422.

Long, H., & Pang, W. (2015). Rater effects in creativity assessment: A mixed methods investigation. *Thinking Skills and Creativity, 15*, 13-25. https://doi.org/10.1016/j.tsc.2014.10.004.

Michel, M., Révész, A., Lu, X., Kourtali, N. E., Lee, M., & Borges, L. (2020). Investigating L2 writing processes across independent and integrated tasks: A mixed-methods study. *Second Language Research, 36*(3), 307-334. https://doi.org/10.1177/0267658320915501.

Myford, C., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement. *Journal of Applied Measurement, 5*(2), 189-223.

Nikmad, F., & Tavassoli, K. (in press). The impact of test length on raters' mental processes while scoring test-takers' writing performance. *Journal of Language Horizons*. https://doi.org/10.22051/LGHOR.2022.37340.1545

Plakans, L. (2010). Independent vs. integrated writing tasks: A comparison of task representation. *TESOL Quarterly, 44*(1), 185-195. https://doi.org/10.5054/TQ.2010.215251.

Pourdana, N., Nour, P., & Yousefi, F. (2021). Investigating metalinguistic written corrective feedback focused on EFL learners' discourse markers accuracy in mobile-mediated context. *Asian-Pacific Journal of Second and Foreign Language Education, 6*(7). https://doi.org/10.1186/s40862-021-00111-8.

Ruiz-Funes, M. (2001). Task representation in foreign language reading-to-write. *Foreign Language Annals, 34*(3), 226-234. https://doi.org/10.1111/j.1944-9720.2001.tb02404.x.

Şahan, Ö., & Razı, S. (2020). Do experience and text quality matter for raters' decision-making behaviors? *Language Testing, 37*(3), 311-332. https://doi.org/10.1177/0265532219900228.

Shi, B., Huang, L., & Lu, X. (2020). Effect of prompt type on test-takers' writing performance and writing strategy use in the continuation task. *Language Testing, 37*(3), 361-388. https://doi.org/10.1177/0265532220911626.

Suskie, L. (2008). Using assessment results to inform teaching practice and promote lasting learning. In G. Joughin (Ed.) *Assessment, learning and judgment in higher education* (pp. 1-20). Springer Science & Business Media. https://doi.org/10.1007/978-1-4020-8905-3_8.

Swales, J. M., & Feak, C. B. (2004). *Academic writing for graduate students: Essential tasks and skills* (vol. 1). University of Michigan Press Ann Arbor, MI. https://doi.org/10.3998/mpub.2173936.

Uludag, P., McDonough, K., & Payant, C. (2021). Does prewriting planning positively impact English L2 students' integrated writing performance? *Canadian Journal of Applied Linguistics, 24*(3), 166-185. https://doi.org/10.37213/cjal.2021.31313.

Van Moere, A. (2014). Raters and ratings. In A. Kunnan (Ed.), *The companion to language assessment, Vol III: Evaluation, methodology, and interdisciplinary themes* (pp. 1358-1374). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118411360.wbcla106.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287. https://doi.org/10.1177/026553229801500205.

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

Wolfersberger, M. A. (2007). *Second language writing from sources: An ethnographic study of an argument essay task* [Unpublished doctoral dissertation]. University of Auckland.

Zabihi, R., Mehrani-Rad, M., & Khodi, A. (2019). Assessment of authorial voice strength in L2 argumentative written task performances: Contributions of voice components to text quality. *Journal of Writing Research, 11*(2), 331-352. https://doi.org/10.17239/jowr-2019.11.02.04.

Zanders, C. J., & Wilson, E. (2019). Holistic, local, and process-oriented: What makes the University Utah's Writing Placement Exam work. *Assessing Writing, 41*, 84-87. https://doi.org/10.1016/j.asw.2019.06.003.