ATU PRESS

ISSUES IN LANGUAGE TEACHING (ILT)
VOL. 13, NO. 1, 183-218, June 2024
https://doi.org/10.22054/ilt.2024.76325.812 Document Type: Research Paper

Iranian EFL Writing Raters' Behaviors When Interacting with Analytic and Holistic Rating Scales

Narjes Khodaparast
MA in ELT, Persian Gulf University, Bushehr, Iran

Nasim Ghanbari*
Assistant Professor of ELT, Persian Gulf University, Bushehr, Iran

Abbas Abbasi
Assistant Professor of ELT, Persian Gulf University, Bushehr, Iran

Received: October 17, 2023; **Accepted:** June 30, 2024

Abstract

Among different factors affecting writing assessment, rater and rating scale are two influential variables that determine the outcome of assessment. In fact, the way raters interact with rating scales considerably affects the validity of their assessment. Taking this into account, this study attempted to identify and classify the raters' behaviors in the Iranian EFL context when using analytic and holistic rating scales. To this end, a group of nine expert raters were asked to verbalize their thoughts when rating student essays. They were also asked to do their rating using the analytic scale of ESL Composition Profile and IELTS holistic scale. Upon the qualitative analysis of think-aloud protocols (TAPs), two themes emerged which showed the raters' behaviors when applying the rating scales. The findings further showed that when using the holistic scale, the raters read the text first to get an overall impression. Then they assessed the text based on their own criteria. Next, they referred to the scale for scoring and in the last stage they provided evidence for their scoring. On the other hand, when applying analytic rating scales, the raters first scanned the text for surface features. Then they read the text for their initial impression. Next, they read each scale component and its descriptor for scoring and finally, they attempted to provide evidence for their scoring. In addition to identifying the raters' behaviors, the raters' behaviors were classified to shed light on the process of rating. The findings imply that the diagnosis of the rater-rating scale interactions can unveil the strengths and weaknesses of the EFL rating process. This, in turn, can provide more quality training for the raters and in this way enhance the scoring validity of their judgements. In the long run, this would improve the professional development of the raters involved in the writing assessment.

Keywords: rater, rating scale, holistic scale, analytic scale, writing assessment, think-aloud protocols (TAPs)

*** Corresponding author's email:** btghanbari@pgu.ac.ir

INTRODUCTION

The reliability of performance assessment has been very problematic and a cause of concern for a long time. Outcomes significantly depend on the raters, so rater-related factors considerably affect the evaluation of writing. According to Myford and Wolfe (2003), some of the criteria associated with rater effects include leniency/severity, central tendency, randomness, halo effect, and differential leniency/severity. Using rating scales by trained raters can help produce more valid and reliable results. Therefore, the existence of a reliable and valid rating scale is necessary for assessing writing. So, rater and rating scale are two influential components of writing assessment that interact with each other to reach a final outcome. Although many studies (Alotibi & Alshakhi, 2022; Barkaoui, 2010c; Heidari et al., 2022) have investigated the rating process, it still faces many difficulties and the process of rating is unclear.

The role of rating scales and raters' behaviors as a source of error during the rating process has been the focus of many scholars (Alotibi & Alshakhi, 2022; Ayoobiyan & Ahmadi, 2023; Barkaoui, 2010c; Barkaoui, 2011; Jia & Zhang, 2023; Shin et al., 2023). Cumming (1990) stated that little is known about how composition evaluation is developed, and the thinking processes and knowledge that are needed for evaluation are vague. In fact, without knowing what is going on inside the raters' minds we cannot assert if their evaluation is fair (Cumming et al., 2001, 2002; Sakyi, 2000). Along the same lines, Connor-Linton (1995) stated "if we do not know what raters are doing and why they are doing it, then we do not know what their ratings mean" (p. 763). It is hard to judge the fairness of the raters' evaluations if there is no clear picture of what is happening in the their cognitive processes (Cumming, 2002; Sakyi, 2000).

The nature of writing assessment is multifaceted. A human rater performs the rating task within many contextual factors. In fact, the score assigned to a text is not just based on the writing quality. Rather, the outcome is affected by a myriad of factors and the interaction between them, such as

rater, test-taker, scale, scoring, and practice. As a result, diverse interactions with each of these assessment components can cause different outcomes (Lumley, 2005).

Moreover, different raters have different criteria and interpretations, so disparate results can be achieved from various raters. In fact, rating scales are applied to reduce these discrepancies. Many researchers such as McNamara (1996) mentioned many factors that affect performance assessment among which rating scales are of great importance to enhance the quality of rating. Moreover, Nakamura (2004) examined the advantages and weaknesses of rating methods and claimed that having a clear-cut rating scale can greatly help.

In addition to rating scales, many studies have examined the rating process (Cumming, 1990; Cumming et al., 2001; Jia & Zhang, 2023; Lumley, 2000; Milanovic et al., 1996; Shin et al., 2023; Vaughan, 1991). They consider many different factors that affect the rating process. Among these, raters and rating scales are considered influential factors that considerably affect writing assessment (McNamara, 1996). Therefore, raters' behavior toward rating scale has been investigated in many studies (Abedi, 2010; Cumming, 1990; Han & Huang, 2017; Kim & Lee, 2015; Lumley, 2005; Polat, 2020). Although numerous studies have examined the rater behavior during the rating process, more research is needed to investigate raters' behaviors when using the scale. This is because past studies have mostly focused on the raters and rating scales separately (Barkaoui, 2007; Eckes, 2008; Erdosy, 2004; Kim, 2015; Polat, 2020; Şahan & Razı, 2020). Moreover, many studies have adopted a quantitative approach to investigate the rater variability in writing assessment tasks (Ayoobiyan & Ahmadi, 2023). Although numerical outcomes of these studies are helpful, they still fail to provide a detailed account of how the raters do the rating task using the rating scales. The analysis of rating behaviors would provide a robust basis for the validation of the scores assigned to the texts. In fact, in the absence of rater training programs, many raters draw on their expertise to justify their scoring (Ghanbari & Barati, 2020). Examining the raters' minds when

interacting with the scale dimensions would help elicit their particular cognitions when doing the rating. As a result, the validity of the rater's scoring judgment would be determined. Hence, in the absence of such exploratory studies, the present study aimed to elicit and classify the rating behaviors of Iranian EFL raters using holistic and analytic rating scales.

LITERATURE REVIEW

Rater in Writing Assessment

In performance assessment, human raters significantly affect the outcome, while in objective assessments only the test-taker and the task are effective factors (McNamara, 1996). Therefore, raters have been at the center of attention in recent research. Barkaoui (2010b) asserts that raters are the most notable sources of variability in terms of the scores assigned and the frequency of the decision-making strategies used for rating.

Rater-related factors greatly affect the evaluation of writing. Researchers have counted several factors. Myford and Wolfe (2003) proposed leniency/ severity, central tendency, randomness, and halo effect. In addition, some factors like raters' academic and linguistic background, professional experience, native language, rater training, written script, rating methods, rating criteria, tolerance for errors, and rating scales also affect the writing assessment (Ahmadi, 2019; Barkaoui, 2010a; Bijani & Said Bani Orabah, 2022; Eckes, 2008; Huang, 2012; Khalilzadeh & Khodi, 2021; Lim, 2009; Lumley, 2002). Moreover, the sequencing and wording of the rating scales' components and their idiosyncratic rating style have been shown to impact the rating process. Style means the way the raters do their rating and assign scores based on the text, rating scales, and their interpretations (Lumley, 2002; Sakyi, 2003; Smith, 2008). It needs to take into account the rating procedure adopted by the raters. Knoch (2011) defines rater behavior as the stages the raters follow during the rating process, the strategies they use, and the aspects of writing they concentrate on. Following McNamara (2000), the score assigned to a text is the outcome of the rater's behaviors and judgment

as well as the writer. As a result, not only is the role of test taker important, but the raters and their behaviors during the scoring procedure also have a significant effect.

Raters play a central role in writing assessment due to their different background, experience, expectations, rating process, etc. (Huang, 2009; Weigle, 2002). Raters may have a unique way of thinking about discrete features of writing (Şahan & Razı, 2020). Therefore, knowing the raters' thoughts and behaviors during the rating process is important. At the end of the rating process, different raters may come to different scores which is due to their distinct behaviors that cause some errors. Therefore, using rating scales may help raters reach some agreement when rating and assigning final marks.

Rating Scale in Writing Assessment

Rating scale is the essential part of assessment. There are various kinds of rating scales with different designs and various assessment factors. Rating scales consist of well-established criteria and scores that are assigned to a spoken or written performance (Campbell et al., 2000). They are "realizations of theoretical constructs and beliefs about what writing is and what matters about writing" (Hamp-Lyons, 2011, p. 3).

Quality of writing assessment remarkably depends on the criteria of the rating scale used during the rating (Ghanbari et al., 2012). According to Lumley (2002), the main purpose of using rating scales is to help the raters do their ratings more reliably, systematically, and in a categorized way. Using the rating scales results in consistent scores. In addition, Knoch (2009a) mentioned that inexplicit descriptors do not provide satisfactory guidelines, so raters may fall back on their own impressions of the text, which could decrease reliability. On the other hand, Smith (2000) stated that detailed and precise descriptors could increase reliability. Based on these two studies, it can be concluded that the descriptors of the rating scales are also of great importance**.** However, Rezaei and Lovorn (2010) found that the effect of

using rating scales is not significant and may not improve the reliability and validity of assessment unless the raters receive adequate pre-rating training.

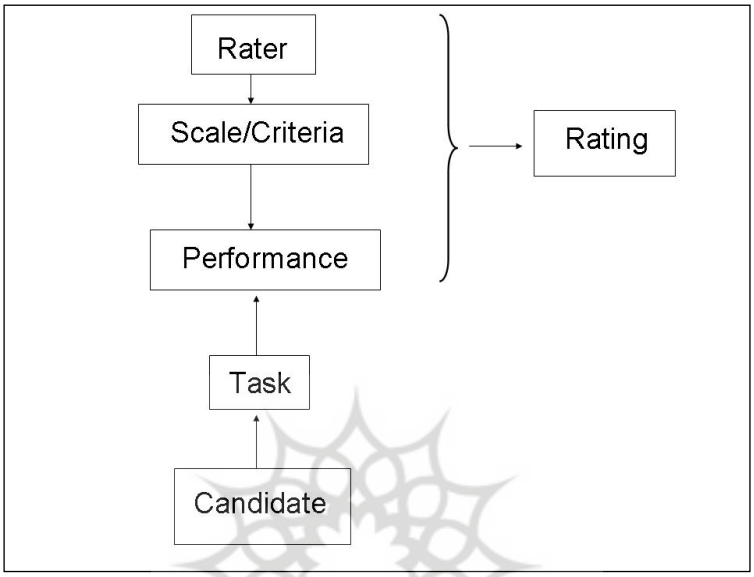

Figure 1: Factors influencing writing performance (McNamara, 1996)

Rater and rating scale are not separate entities, rather they are interconnected in the rating process (Ghanbari & Barati, 2014). A number of qualitative studies have been conducted to gather evidence on raters' interpretation and practical use of both the old and the revised IELTS rating scales (Shaw & Falvey, 2008). Li and He (2015) endorsed McNamara (1996), who explained that the interaction between rater and scale mediates scoring of performance tests. However, they have identified different interaction patterns. In this regard, Barkaoui (2007) concluded that his findings suggested significant interaction effects between scales and raters. Furthermore, Barkaoui (2008, 2010b) found that rating scales greatly influence raters' rating behaviors, even more than their rating experience. His studies revealed substantial evidence concerning the impact of rater-scale interactions during the rating process.

Rater-Rating Scale in Writing Assessment

According to a number of studies conducted on writing assessment, raters and rating scales are a matter of concern in the EFL context. Raters play a critical role in writing assessment. Studies have shown that even experienced raters face challenges when assessing texts (Wolfe et al., 2016). Therefore, some studies have been conducted to show how rating scales affect the outcome of the rating process (Barkaoui, 2007; Sakyi, 2000; Vaughan, 1991). Also, some researchers have compared holistic and analytic rating scales (Alotibi & Alshakhi, 2022; Bacha, 2001; Knoch, 2009b; Nakamura, 2004). Rating scales can affect raters' rating processes (Barkaoui, 2010a), but few studies have explored the processes raters go through while interacting with the rating scales. Ayoobiyan and Ahmadi (2023) investigated the rater-rating scale interaction in terms of the possible halo effect in their rating. Five raters who had received specialized training applied a four-criterion rating rubric to analytically rate texts on two argumentative topics. The results showed that, except for one rater, the raters did not exhibit any sign of the halo effect across the rubric criteria. The study emphasizes the importance of training in minimizing rater variance when interacting with rating scales in EFL writing assessment.

In another study, Shin et al., (2023) examined the effect of the mode of scoring on raters' rating behavior. A group of six raters were asked to rate both handwritten and typed versions of 82 argumentative essays. The results of *FACETS analysis* showed that mode of scoring affected rater behavior. Adopting a problem-solving approach, Jia and Zhang (2023) explored the raters' scoring behaviors in an integrated writing task. For this aim, a group of six expert raters were asked to verbalize their thoughts while rating the essays. The analysis of the *think-aloud protocols* showed that the raters went through two stages: building text images as isolated nodes and building holistic text images for each dimension, as two sub-goals, respectively. In order to achieve the first sub-goal, raters used strategies such as single-focus evaluating, diagnosing, and comparing; for the second sub-goal, they mainly

used synthesizing and comparing. Moreover, the results showed that the raters resorted to two groups of strategies: demarcating boundaries between scores within a dimension and discriminating between dimensions, each group consisting of more specific processes.

Overall, it can be inferred from the above studies that the evaluation of writing is a subjective process, and without having a clear picture of what happens in raters' minds, it is difficult to claim whether their rating outcomes are valid or not (Cumming et al., 2002; Ghanbari & Barati, 2014). Therefore, scoring validity and consequential validity would be at risk (Weir, 2005).

As long as assessment and rating are conducted by humans, they are inevitably influenced by personal beliefs, ideas, and perceptions (Eckes, 2008; Heidari et al., 2022). The dimensions of the rating scale, the wording of the rubric, the impressions and subjectivities of the raters, and the potential interaction of these elements can influence the raters' evaluations and, therefore, the scores they assign to the texts (Goodwin, 2016). Despite the prominence of the rater's actions, few studies have investigated the black box of rater–rating scale in a detailed qualitative way. Consequently, there is a gap in the literature regarding the identification and classification of the raters' behaviors when using and interacting with the rating scale in the Iranian EFL context. The present study was an attempt to fill this gap in the existing literature.

PURPOSE OF THE STUDY

Many researchers have attempted to explore raters' behaviors (Knoch, 2011; Li & He, 2015; Lumley, 2002; Ostovar, 2011), the effect of rating scales on writing assessment (Ghanbari et al., 2012) and the interaction between raters and rating scales (Barkaoui, 2007, 2008, 2010c; Li & He, 2015). Some factors, such as rating procedures and rating behaviors, particularly among different possible rating stages have received little attention regarding the stages raters go through when using the scale in the Iranian EFL context.

Considering the above, the present study was conducted to address the following two research questions:

1. What are the Iranian EFL writing raters' behaviors when using holistic and analytic rating scales?
2. How can Iranian EFL writing raters' behaviors be classified when using the rating scale?

METHOD

Design of the Study

The design of this study was qualitative, with the researchers focusing on the raters' rating processes using *think-aloud protocols (TAPs)*. Because the rating procedure is a complex task, adopting a qualitative and exploratory design can be helpful (Creswell & Poth, 2018). As a result, the researchers can more effectively interpret the depth, diversity, and complexity of raters' thought processes.

Participants

Participants were nine experienced Iranian EFL university professors from Persian Gulf University (PGU) and Salman Farsi University of Kazerun (KSFU). They were recruited through a convenience sampling procedure. All participants had extensive experience in teaching and assessing writing. As shown in Table 1, all participants had at least five years of experience teaching and assessing writing. The group included five male and four female raters whose ages ranged from 31 to 59 years. Four male raters and one female rater held M.A. degrees, and the remaining raters held Ph.D. degrees. Additionally, three of the raters had attended workshops on writing assessment. To maintain anonymity, the raters are referred to as R1, R2, …, and so on throughout the study.

Table 1: Demographic information of the participants in the study

Raters	Gender	Age	Education	Teaching experience	Assessing writing experience	Workshop participaton rating
R1	Male	59	PhD	Over 20	14-19	No
R2	Female	42	PhD	14-19	14-19	Yes
R3	Female	40	PhD	9-13	3-8	No
R4	M ale	35	MA	14-19	14-19	No
R5	Male	49	MA	Over 20	14-19	Yes
R6	Male	38	MA	9-13	3-8	Yes
R7	Male	55	MA	3-8	3-8	No
R8	Female	42	MA	Over 20	3-8	No
R9	Female	38	PhD	14-19	9-13	No

Instruments

Writing Task

The writing sample used in the current study was randomly selected from a collection of 30 essays written by undergraduate students at Persian Gulf University. After completing their fourth semester, the EFL students were asked to write an expository essay (approximately 160 words) on the following topic: *Recovering emotionally from a disaster* (see appendix A). The texts were written under exam-like conditions. The essays collected in this way were used for two rounds of analytic and holistic rating. The raters were asked to assign a score from 1 to 100 for the analytic scale and from 1 to 9 for the holistic scale.

Think-Aloud Protocols (TAPs)

To gain access to the raters' rating processes, TAPs were used since they reveal the way raters reach a particular score. TAPs elicited information about the raters' performance when they were engaged in the rating task (Huot, 1993). TAPs require that the participants verbalize their thoughts when performing a specific task. This allowed the researchers to directly observe

the processes involved in rating. The TAPs obtained were then analyzed to infer the raters' rating processes.

Rating Scales

In this study, the IELTS holistic scale, which is used to assess the test writing section of the IELTS test, and the analytic rating scale of the ESL Composition Profile (Jacobs et al., 1981) were used. The holistic scale scores essays through a simple grading structure that evaluates a paper's overall quality rather than its individual components. The IELTS holistic scale used in this study consisted of nine bands ranging from *expert user* to *did not attempt the test*.

Additionally, in this study, the analytic rating scale (Jacob's et al., 1981) was used. This scale is widely used for evaluating writing in EFL/ESL contexts. The scale is divided into five main writing components. Each component and level assesses different aspects of writing and clearly describes the writer's performance. The components include content and organization which refer to topic development, unity, cohesion, and coherence. Vocabulary as the second component refers to the range, effective word choice, appropriate register, and word form mastery. The third component of language use focuses on grammatical points, complex structures, accuracy, and syntax. The last one is mechanics which is concerned with paragraphing, capitalization, and superficial aspects of a text. Each component is rated based on four levels, from very poor to very excellent. The score differs for each level in each component.

Training Manual

In this study, the researchers provided a training manual which clarified the think-aloud procedure for the rater participants. In addition to English instructions, the researcher also explained the processes in Persian (native language of the participants).

Demographic Information Sheet

In order to collect information about the raters' personal characteristics, the researcher asked them to fill a personal demographic information form. It asked for their name, age, gender, education, marital status, years of teaching and writing assessment, and participation in any workshop related to writing scoring.

Data Collection Procedure

Before beginning the data collection, the researchers had arranged meetings with the participants. Prior to TAPs sessions, each rater completed a demographic information form asking for their name, age, gender, education, marital status, years of English and writing teaching, years of writing assessment experience, and participation in any workshop related to writing scoring.

Data collection procedure was conducted in two phases: First, all the raters received an initial training and detailed description of the think-aloud procedure through the audio file. The raters were asked to verbalize their thoughts while rating the texts. They were asked to use a talking sign (e.g., a ☆) to remind themselves to verbalize all their thoughts. Next, the raters received two holistic and analytic rating scales. They were asked to rate a text twice, once using Jacob's et al (1981) rating scale, which is a well-known analytic scale, and then by using the IELTS holistic scale. While rating the essay, the raters verbalized their thoughts.

Data Analysis

To analyze the TAPs, qualitative content analysis was used. In this study, conventional model of qualitative content analysis was used (Hsieh & Shannon, 2005). As in conventional content analysis, coding categories were directly derived from the text data. First, the recordings were transformed into textual forms and full transcripts of the recordings were prepared. Because

the raters were allowed to use either Persian or English during the think-aloud sessions, some texts also needed to be translated into English.

The first stage of analysis was pre-coding which included reading and re-reading the transcripts. The researcher read the transcripts many times to shape her thoughts about the data and look for ideas that led her to specific themes. Then, she started reading the texts from the beginning and highlighted any relevant part to the topic and added an informative label. Even if it was not directly related to the focused area, any relevant part was highlighted. Descriptive codes were used in this phase, which were later substituted by pattern codes in the following sessions. Any special feature of the data were highlighted in order to be linked under broader topics.

Next, all the identified codes were listed (Appendices I & II). There were some similar and related codes which were clustered under a broader code. The newly shaped broad categories were checked to see if the new code can be applied to all of them. After revising the list of codes, the researcher returned to the original transcripts and recoded them based on the new categories. After finalizing the revised codes, the researcher considered all the final codes, found a theme for the raters' rating behaviors using each scale, and then classified the emerged behaviors.

RESULTS

Raters' Rating Behaviors When Using the Holistic Scale

The analysis of TAPs revealed how raters behaved when using the holistic rating scale. As Table 2 below shows, the raters' behaviors were categorized into four major ones. The following sections explain the behaviors in detail.

Table 2: Raters' rating behaviors when using the holistic scale

Initial reading of the text to get an overall impression
Assessing the text based on their own criteria
Referring to scale for scoring
Providing evidence for their scoring

Initial Reading of the Text to Get an Overall Impression

Raters looked at the text and expressed their first impression by paying attention to the appearance, handwriting, number of paragraphs, length, title, etc. Then, they continued with reading the whole text. As evidence, rater 2 stated the following:

R2: First, I look at the text. The first thing I noticed was her handwriting which may cause me problem to read it. Also, the cross out in the last line had a negative impression on me. I insist on reading the whole text first.

Assessing the Text Based on Their Own Criteria

After reading the texts, raters started to assess them and comment on different parts like vocabulary, grammar, mechanics, and so on, based on their own criteria of rating and without considering the scale criteria. They commented on different aspects of the texts. The following excerpts show some aspects they attended to when doing their rating:

R1: This paragraph seems not to have a topic sentence, or if it does, the following sentences do not develop the topic sentence.
R9: when you have a body paragraph you should have the topic sentence and then you should have some major and minor supporting sentences.
R5: instead of comma I think there should be an and! Also, all the sentences are simple or compound, I cannot see any complex sentences!
R7: She has good control over her language but she is not, you know, an expert in using the language because again I see that the sentences are not collocative enough.

Referring to Scale for Scoring

After their initial assessment of the text, the raters referred to the scale and scanned it to give their score. The raters had their own approach when referring to the scale for scoring. The following excerpts show how they used the scale for rating.

R2: Expert user, 9, description is 'The test taker has fully operational command of the language. Their use of English is appropriate, accurate and fluent, and shows complete understanding'. The writing was not like this, so, I look at very good user very quickly, also good user, competent user and modest user. Let's look at modest user and limited user deeply. 'The test taker has a partial command of the language and copes with overall meaning in most situations, although is likely to make many mistakes. He/she should be able to handle basic communication in own field'. I read extremely limited user and compare it with the one before it. Limited user is better describing her. I give her 4.

Rater 7 read all the bands and the descriptors and commented if this band describes the writer. Some raters just went straight to the one or two which they thought were best describing the writer's level to see if it is ok or not. The following excerpt from rater 1 shows how he approached the scale:

R1: Regarding our scale of scoring, I think, uhm…. This text's score falls in…. uhm…6 band. 'The test taker has an effective command of the language despite some inaccuracies, inappropriate usage and misunderstandings. He/she can use and understand fairly complex language, particularly in familiar situations' No, I think I give her 7. Yes 7.

Rater 3 directly referred to the band score that she thought was the best:

R3: According to what I've read and the points that I mentioned, this writer gets 6, competent user, 'The test taker has an effective command of the language despite some inaccuracies, inappropriate usage and misunderstandings. He/she can use and understand fairly complex language' based on this topic, she showed a partial complexity but not the acceptable one. According to the descriptions which are in my mind she is not expert user, very good user, or good user.

Rater 8 did not read all the bands. Rather, she read only some parts of band 5 description and considered it as less than appropriate (3.5 or 4). Then, she provided some evidence which made her change her scoring to 4.5.

R8: based on holistic view of IELTS, this writing cannot get more than 5. Now, I read the criteria of the scale, 'The test taker has a partial command of the language and copes with overall meaning'. In my view, this writer is 3.5 or 4. She has breakdowns in communication and basic competence is limited to familiar situations.

Providing Evidence for the Scoring

After agreeing on the score, the raters provided some evidence to validate their scoring. The following excerpts by raters 2 and 9 show how they justified their particular scoring.

R2: I think limited user is better because she had many problems in different parts of organization and she actually couldn't use complex language. Also, her chosen vocabularies were very simple.

R9: I think modest user is good, of course I think that she doesn't ..[pause].. know anything about the mechanics, organization of writing…this is not a piece of writing for an English student actually should write...[pause]... Just she tries to write some sentences after each other without any good organization so not a good introduction, not a good body, not a good conclusion" or
rater 6: "this writing is between good user and competent user, here are the reasons: the first one is that she partially understands the topic but in case of word complexity and collocations sometimes she is a good user and sometimes competent user. I do not see any grammatical problems but in case of cohesion and coherence there are some discrepancies. In the case of vocabulary, she can use better synonyms. She does not develop the topic very well and does not answer the question properly.

Raters' Rating Behaviors When Using the Analytic Scale

Table 3 below shows the rating behaviors of the raters when using the analytic rating scale. Each of these behaviors will be elaborated upon in the next sections.

Table 3: Raters' rating behaviors when using the analytic scale

Scanning the text for surface features
Reading the text for an initial impression
Reading each component and its descriptors for scoring
Providing evidence for the scoring

Scanning the Text for Surface Features

Raters started the rating process by briefly looking at the text to get an overall initial impression. The raters considered handwriting, neatness, number of paragraphs, clarity, etc. The following excerpts show how the raters considered different superficial aspects of the texts in their first reviews of the texts.

R2: first, I look at the text. The first thing I noticed was her handwriting which may cause me problem to read it. Also, the cross out in the last line had a negative impression on me. Now, I have a general impression in my mind.
R9: let me have a look at the text, it is arranged into 3 paragraphs, so it seems it is an essay.
R8. Regarding the handwriting I think this person is not at a very high level. I see that's just a few paragraphs, it's not going to be enough because at least 4 paragraphs should be included.

Reading the text for an initial impression

After having a quick look at the text, the raters read the whole text to have a passing acquaintance with the text.
R2: first I look at the text, I have a general impression in my mind, now I am going to read the whole text.
R7: first, I read the whole text several times
R 3: let me have a look at the text.

Reading each component and its descriptors for scoring

After reading the text and getting familiar with it completely, the raters referred to scale and read each component and its descriptor.
Rater 2 read the components and its criteria but not all of them, mostly the highest one:

R2: let's shift to scale. Content: the definition is 'excellent to very good: knowledgeable, substantive, thorough development of thesis, relevant to assigned topic' absolutely it was not like this, I can give her 13-16. 'Very poor: does not show knowledge of subject, non-substantive'. Now between 13 and 16, I choose 14. Because it really has a weak content. Let's go to the next component, organization. It means: 'fluent expression, ideas clearly stated or supported, well-

organized, logical sequencing' unfortunately it was not like this, so I think very poor is ok. I give her 8. Vocabulary part, Jacob's scale says: 'sophisticated range, effective word idiom choice and usage, word form mastery, appropriate register' I do not give her very poor but fair to poor is ok. I give her 12, the middle range". For the next two components she also read the highest criteria and assigned score.

After reading the whole text, rater 1 read it again in detail for the components of the analytic scale. Focusing on different parts, he went through the scale:

R1: I think content falls between 21-17 because the development of ideas is inadequate, so 21 to 17. In terms of organization can fall fair to poor, non-fluent, ideas confused or disconnected, although the paragraphs are not well-developed, the way has written sentences ...[pause]... . And ideas are linked, falls 17 to 14 organization. Vocabulary: words are repeated several times, for example control. I think it falls within fair to poor. And language usage: sentences are not complex, rather, they are either compound or simple sentences. It falls between 17 to 11 and mechanics, occasional errors of spelling, punctuation, capitalization, I'll give it 5. So, if I add these scores…um… 21 plus 17 plus 13 and plus 17 and 5...umm... If I add them up 8…11…18...23... 4, 5… 73... I will give it 73". He did not read all the criteria, just the components and the levels and assigned the score.

Rater 4 just read the components and the levels and chose the score:

R4: *So, if I want to grade this text based on the Jacob's scale for its content, this is far from good, I think it's very poor which I give it 13 to 16 out of 30. In terms of organizational, I think it's, you know, when the content is not relevant other things lose meaning, but ok, in terms of what she has written, thinking that the text topic was something*

else. The organization would be fair to poor again. Fair to poor Last one was very poor this one is fair to poor, that would give it 10 to 13 out of 20. In terms of vocabulary, I would give it either fair to poor or good to average, I think good to average, that's not to be too strict, so that would be 14 to 17 out of 20. And in terms of language use I would give her fair to poor, which is 11 to 17. And in terms of mechanics, it would be fair to poor or good to average that would be 3 or 4 out of 5".

Providing Evidence for Scoring

Most of the raters provided evidence in order to validate their scores. They read each component then provided evidence after scoring it. For example, rater 2, in order to prove her scoring of organization part, said:

R2: She couldn't define the thesis sentence in the body, she talked about 'some ways' in her thesis but she wrote just one paragraph in the body which is completely offtopic as well. So, I gave her 8.

Rater 3 after reading mechanics part, provided these evidences for the score she gave to this part:

R3: she used punctuation marks. Her spelling and paragraphing were good; however, it could be better. She used comma after 'however'. Looking at the second paragraph, it is clear that she used all the punctuations so I gave her excellent to very good.

Rater 6, reading organization part, stated: *considering organization she gets 'good to average'. Why not 'fair to poor'? She has pretty good organization but considering 'cohesion', this writing is medium in using transitional devices. It has poor 'referencing' although we can understand the meaning.*

Topic related vocabulary is not excellent, also repetition occurs a lot which affects the organization part and cannot be very good.

Rater 8 validated her scoring to content part by stating *I give her 17-21 because she knows the subject and some related vocabularies like 'catastrophe' but she did not address the topic and did not develop the ideas.*

Classifying the Raters' Behaviors

The second research question aimed to classify the raters' behaviors when involved in the rating task. Table 4 below shows the classification of the raters' behaviors into four main steps.

Table 4: Classification of the raters' behaviors

Initial interpretation
Self-monitoring judgment
Scale-monitoring judgment
Inferential judgments

Initial Interpretation

This stage included reading the text and the scale and articulating a general impression. Getting started with the rating task, the most frequently observed rating behavior was "looking at the appearance of the text or reading the whole text to get an overall first impression".

R1: I will read the text first and then I will have my overall assessment of the text. The essay is a bit short. It has no title.
R2: First, I look at the text. The first thing I noticed was her handwriting which may cause me problem to read it. Also, the cross out in the last line had a negative impression on me. I insist on reading the whole text first.

Raters were supposed to rate the texts two times by using holistic and analytic scales, so they mentioned which scale they were going to use in advance. For example:

R7: first, let's consider analytic scale which I prefer it to holistic because it has different parts.

R2: Scoring the article based on the holistic method. I've read the text first.

Two raters read the scales carefully before rating:

Rater 2 said: *I looked at the scale, reviewed it, read the components like content, vocabulary, language use and so on and all the criteria.*

Rater 7 also mentioned: *I carefully read the scales several times to get their points.*

Self-monitoring Judgment

The raters read the text and did their initial assessment by using their own criteria, commenting on different parts; looking for thesis statement, anchoring, predicting, repeating, and so on. They also commented on the writer's performance level.

R8: The writer performed very poorly!

R5: It seems that the student has not passed some courses on how to write outline.

R1: I will have my overall assessment of the text.

R2: I'm looking for the thesis, because it is an essay, but it does not seem to have a thesis.

R8: From the first sentence, I can understand that the writer has a poor command of English.

R4: *The level of vocabulary is not as professional as I expect of a pro-user to be*

Scale-monitoring Judgment

Referring to the scale, the raters read the components and the descriptors. They analyzed the text based on the components; however, they did not read all the criteria. In analytic rating, most of the raters did not read the high-level criteria. On the other hand, in holistic rating while some raters read all the band scores, some just read the ones that they considered as appropriate. The raters referred to the text to check the components several times.

R4: Oh, I think, you know, 5 or 6 out of 9. Modest user or competent user, perhaps 5.5 would best describe her.
R3: I return to text to check the vocabulary and start reading vocabularies up and down.
R8: I didn't pay attention to capitalization. Let me check… now I see she follows the capitalization rules.

Inferential Judgments

After referring to the scale, the raters assigned their scores and provided evidence to validate their scoring.

R6: This text can be placed between good user and competent user. Here are the reasons: the first one is that she partially understands the topic, but in case of word complexity and collocations sometimes she is a good user and sometimes competent user. I do not see any grammatical problems but in case of cohesion and coherence there are some discrepancies. In the case of vocabulary, she can use better synonyms. She does not indicate the topic very well and does not answer the question properly.
R2: I think limited user is better because she had many problems in different parts of organization and she actually couldn't use complex language. Also, her chosen vocabularies were very simple.

R8: I give her 17-21 because she knows the subject and some related vocabularies like 'catastrophe' but she did not address the topic at all and did not develop the ideas.

Two raters who predicted or assigned their scores before referring to the scale, revised their scores after that. Adopting the analytic scale, some raters summed up the scores and some did not.

DISCUSSION

Many studies have investigated the rater and rating scale while assessing writing, mostly to improve reliability and validity (East, 2009; Eckes, 2008; Johnson & Lim, 2009). This study was an attempt to look into the writing raters' behavior when using the rating scales. The analyses of the TAPs revealed different rating behaviors of the raters while using the rating scales. Different models have been proposed. For example, Cumming (2002) proposed a descriptive framework of the rater's decision-making behaviors during the evaluation of EFL/ESL writing task. McNamara (1996) also stated factors influencing writing performance. Ghanbari and Barati (2014) also presented the stages of the rating process. This study proposed two themes which showed the kind of raters' behaviors when using a holistic or analytic scale.

Considering the first question, it was found that the raters had an initial assessment and then they conducted their self-monitoring assessment. Many studies consider the rater at the heart of the rating process (Cumming et al., 2001; Erdosy, 2004; Ghanbari & Barati, 2014; Lumley, 2005). Ghanbari et al. (2012) revealed that raters mostly rely on their own criteria when assessing the texts. Similarly, this study found that the raters prioritized their own criteria and did their first assessment based on monitoring themselves rather than referring to the rating scales.

After the raters' initial assessment, they referred to the scale to finalize their scoring. Lumley (2002) claimed that although raters try to follow the

rating scales closely, they are greatly under the influence of their initial reading of the text. Wind (2019) also observed that some raters did not use highest or lowest rating categories of the rating scales. In this study, the influence of the rating scale was observed. While using the scale, rater 2 returned to the text to look for evidence. The scale made her pay attention to a component which she had forgotten to attend to in her initial reading. She praised using the scale. Moreover, rater 4 paid attention to the topic just after returning to the scale. Wiseman (2012) explained that analytic scale might be better than holistic as it is more sensitive to different levels of ability. Rater 7 also stated that he preferred the analytic scale over holistic scale because it has different bands and each band gets its related score. Therefore, based on these scores the text can be assessed. He did not ignore any part of the scale. After reading the descriptors of each component, the raters mentioned the problems of the text. Then, they referred to the scale and assigned the scores. Next, they interpreted their scoring by providing some evidence.

The second research question focused on the classification of the raters' behaviors. The rating process affects the quality of rating**.** Hamp-Lyons (2007) stated that the quality of the rating is greatly considered in the outcome analysis. Upon a careful analysis, the raters' behaviors were classified into four stages. The steps followed by the raters in this study is in line with Cumming (2002) and McNamara (1996). Cumming (2002) presented three stages: first, scanning the composition for surface-level identification; second, engaging in interpretation strategies, reading the essay while exerting certain judgment strategies, and third, articulating a scoring decision while summarizing and reinterpreting judgments. However, because this study focused on the interaction of the rater and rating scale, it only agrees with Cumming's model in some stages. The first stage of Cumming's model is the subset of initial interpretation. The second stage, which is exerting judgment strategies, is in line with the model emerged in this study; however, in this model after self-monitoring judgment, scale monitoring is presented. Also, the last stages are somehow the same. Moreover, as presented in McNamara's model, it can be found that the raters and rating scales have an

important effect on the rating process. Raters focused on their own judgment which is affected by many factors. Scale monitoring judgment also proves the important role of rating scales in the rating process.

CONCLUSION AND IMPLICATIONS

Using TAPs, this study showed that when using a holistic scale, the raters first considered the whole text to examine the surface level. Next, they got engaged in the initial assessment of the text using their own criteria, and then the raters referred to the scale to assign a score, and finally, they provided some evidence to prove their scoring. Using an analytic scale, the raters looked at the text and then read the text to get an overall impression. Next, they referred to the scale, read each component, scored them, and provided evidence for each one. These behaviors have been classified into 4 categories: initial interpretation, self-monitoring judgment, scale-monitoring judgment, and inferential judgment.

In conclusion, the findings of this study showed that Iranian EFL writing raters mostly relied on their own criteria for assessing writing. Especially, when using the holistic scale, they referred to the scale for scoring. The raters had more interaction with the analytic scale than the holistic scale. They considered all the analytic components for scoring.

This study has several implications. The first implication of this study is improving the training process and consequently the writing assessment practice in the EFL context of Iran. Clearly, understanding the rating procedures, training process, and writing assessment can be improved in EFL context. Teachers and raters get familiar with the aspects and components they should assess and practice the use of the rating scales. Training sessions can increase the raters' awareness of their own rating approach and the problems it may have. Raters and writing teachers may receive clear instructions, accompanied by some discussions to identify appropriate and reasonable procedure of rating writings by using the analytic or holistic scale.

The second implication is enhancing EFL raters' professional development in terms of writing assessment behaviors. Rating procedures need theoretical foundations in the EFL context. The third implication would be emphasizing the role of the rating scale in performance assessment and encouraging the right use of the rating scale in the context. The rating scale greatly affects the assessment. The existence of the rating scale improves the consistency of the assessment. In this study, the behaviors of raters regarding the rating scales were examined which would help to signify the place of rating scale among Iranian raters. Next implication is that the raters' interpretations of the scale criteria will provide a better evaluation of the strengths and weaknesses of the students' writing abilities. Therefore, the raters should be familiarized with the scale components and the training sessions are further emphasized.

The last implication would be improving the validity and reliability of the raters' rating behaviors. Raters are among the most influential factors that constrain validity and reliability. This study focused on the raters' behaviors and also their interaction with the rating scale which is a tool for improving validity and reliability. Knowing what they should attend to, they can have more valid and reliable rating. By clarifying the possible stages a rater may go through, the rating process will be generally more consistent. By understanding these aspects, these phases can be used in the best way possible to obtain the highest validity and reliability, to have the least impact on the raters' rating process, and to increase the fairness of the scores.

Future studies can compare a group of ESL raters with EFL raters. It can clarify the dissimilarities between these two groups of raters during the writing assessment and the role of the rating scale in these contexts. A similar study can be done by including the role of gender. The raters' rating behaviors can be compared between the two genders. In this study, few raters participated; therefore, if the study is replicated with more participants, more valid results will be obtained. Future research can also be done with more than one writing sample or use texts of the students at different proficiency levels to see if the raters would follow the same rating process.

This study also faced some limitations. The first limitation was the small number of the participants. Overall, nine raters participated in the study. This is because transcribing, coding, and analyzing TAPs were time-consuming and labor-intensive. In order to generalize the results, a larger number of raters are needed to participate in the study. TAPs were the main data collection instrument in this study. Although the researcher provided the raters with training, raters might not have been completely prepared to verbalize all their thoughts. The use of TAPs influenced both the rating processes and the results (Barkaoui, 2008, 2010a; Lumley, 2005). In other words, the raters' verbalized thoughts might not fully represent their mental processes. Consequently, the limitations of this method can lead to some validity problems. Furthermore, TAPs were coded and analyzed qualitatively which might be the source of some problems. As Barkaoui (2010a) and Cumming et al. (2001) stated, the coding itself has its limitations.

The participants did not receive any special training in using the rating scales. Some researchers proposed that rater variation can be reduced by training the raters (Jacobs et al., 1981; Weigle, 1994). Nevertheless, the fact that many teachers who rate the students' texts do not receive formal training can alleviate this limitation (Elorbany & Huang, 2012; Huang & Foote, 2010).

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Narjes Khodaparast http://orcid.org/0000-0003-1653-1124
Nasim Ghanbari http://orcid.org/0000-0002-1652-8438
Abbas Abbasi http://orcid.org/0000-0004-1673-6543

References

Abedi, J. (2010). *Performance assessments for English language learners*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Ahmadi, A. (2019). A study of raters' behavior in scoring L2 speaking performance: Using rater discussion as a training tool. *Issues in Language Teaching*, *8*(1), 195-224. doi: 10.22054/ilt.2020.49511.461

Alotibi, Sh., & Alshakhi, A. (2022). A comparative study of EFL instructors' essay rating: Holistic versus analytic approaches at a tertiary institution in Saudi Arabia. *Theory and Practice in Language Studies*, *12*(1), 55-64. https://doi.org/10.17507/tpls.1201.07

Ayoobiyan, H., & Ahmadi, A. (2023). Detecting halo effects across rubric criteria in L2 writing assessment: A many-facet rasch analysis. *Applied Research on English Language*, *12*(1), 159-176. https://doi.org/ 10.22108/are.2022.132503.1848

Bacha, N. (2001). Writing evaluation: what can analytic versus holistic essay scoring tell us? *System*, *29*, 371–383. https://doi.org/10.1016/S0346-251X(01)00025-2

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, *12*(2), 86-107. https://doi.org/10.1016/j.asw.2007.07.001

Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL rating outcomes and processes* (Unpublished doctoral dissertation). University of Toronto, Toronto, Canada.

Barkaoui, K. (2010a). 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

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

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education Principles Policy and Practice*, *18*(3), 279–293. https://doi.org/10.1080/0969594X.2010.526585

Bijani, H., & Said Bani Orabah, S. (2022). Facet variability in the light of rater training in measuring oral performance: A multifaceted rasch analysis. *Issues in Language Teaching*, *11*(2), 255-290. doi: 10.22054/ilt.2023.63589.634

Campbell, D. M., Melenyzer, B. J., Nettles, D. H., & Wyman, R. M. Jr. (2000). *Portfolio and performance assessment in teacher education*. Boston: Allyn and Bacon.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, *7*(1), 31–51. https://doi.org/10.1177/026553229000700104

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

Cumming, A., Kantor, R. & Powers, D. E. (2002). Decision-making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal, 86*(1), 67–96. https://doi.org/10.1111/1540-4781.00137

East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing Writing*, *14*(2), 88-115. https://doi.org/10.1016/j.asw.2009.04.001

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

Elorbany, R., & Huang, J. (2012). Examining the impact of rater educational background on ESL writing assessment: A generalizability theory approach. *Language and Communication Quarterly*, *1*(1), 2-24.

Erdosy, M. U. (2004). *Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions*. (TOEFL Research Report RR-03-17). Princeton, NJ: Educational Testing Service.

Ghanbari, B. Barati, H. and Moinzadeh, A. (2012). Rating Scales Revisited: EFL Writing Assessment Context of Iran under Scrutiny. *Language Testing in Asia*, *2* (1), 83-100. https://doi.org/10.1186/2229-0443-2-1-83

Ghanbari, N. & Barati, H. (2014). Iranian EFL Writing Assessment: The Agency of Rater or Rating Scale? *Tabaran Institute of Higher Education*. *4*, 204-228.

Ghanbari, N., & Barati, H. (2020). Development and validation of a rating scale for Iranian EFL academic writing assessment: A mixed-methods study. *Language Testing in Asia*, *10*, 17. https:// doi.org/ 10. 1186/ s40468- 020- 00112-3.

Goodwin, S. (2016). A many-facet Rasch analysis comparing essay rater behavior on an academic English reading/writing test used for two purposes. *Assessing Writing*, *30*, 21–31. https://doi.org/10.1016/j.asw.2016.07.004

Hamp-Lyons, L. (2007). The Impact of Testing Practices on Teaching: Ideologies and alternatives. Cummins, J., Davison, C. (Ed*). International Handbook of English Language Teaching* (pp.487-504). Springer.

Hamp-Lyons, L. (2011). Writing assessment: Shifting Issues, new tools, enduring questions. *Assessing Writing*, *16*(1), 3–5. https://doi.org/10.1016/j.asw.2010.12.001

Han, T., & Huang, J. (2017). Examining the impact of scoring methods on the institutional EFL writing assessment: A Turkish Perspective. *PASAA*,53 January - June 2017

Heidari, N., Ghanbari, N. & Abbasi, A. (2022). Raters' perceptions of rating scales criteria and its effect on the process and outcome of their rating. *Language Testing in Asia 12*(20), 1-19. https://doi.org/10.1186/s40468-022-00168-3

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative health research*, *15*(9), 1277–1288. https://doi.org/10.1177/1049732305276687

Huang, J. (2009). Factors affecting the assessment of ESL students' writing. *International Journal of Applied Educational Studies*, *5*(1), 1–17.

Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment. *Assessing Writing*, *17*(3), 123-139. https://doi.org/10.1016/j.asw.2011.12.003

Huang, J., & Foote, C. J. (2010). Grading between the lines: What really impacts professors‟ holistic evaluation of ESL graduate student writing? *Language Assessment Quarterly*, *7*, 219–333. https://doi.org/10.1080/15434300903540894

Huot, B. (1993). The influence of holistic scoring procedures on reading and rating students' essays. In M. Williamson & B. Huot (Eds.), *Validating holistic scoring for writing assessment* (pp. 206-236). Cresskill, NJ: Hampton Press.

Jacobs, H.L., Zinkgraf, S.A., Wormuth, D.R., Hartel, V.F. *and* Hughey, J.B. (*1981).* Testing ESL composition: a practical approach. *Rowley, MA: Newbury House.*

Jia, W., Zhang, P. (2023). Rater cognitive processes in integrated writing tasks: from the perspective of problem-solving. *Language Testing in Asia*, *13* (50). https://doi.org/10.1186/s40468-023-00265-x

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

Khalilzadeh, S., & Khodi, A. (2021). Teachers' personality traits and students' motivation: A structural equation modeling analysis. *Current Psychology*, *40*(4), 1635-1650. https://doi.org/10.1007/s12144-018-0064-8

Kim, S., & Lee, H. K. (2015). Exploring rater behaviors during a writing assessment discussion. *English teaching*, *70*(1).

Knoch, U. (2009a). *Diagnostic assessment of writing: The development and validation of a rating scale*. Frankfurt: Peter Lang.

Knoch, U. (2009b). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, *26* (2), 275–304. https://doi.org/10.1177/0265532208101008

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing*, *16*(2),81-96. https://doi.org/10.1016/j.asw.2011.02.003.

Li, H., & He, L. (2015) A Comparison of EFL Raters' Essay-Rating Processes Across Two Types of Rating Scales, *Language Assessment Quarterly*, *12*(2), 178-212, https://doi.org/10.1080/15434303.2015.1011738

Lim, G. (2009). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing Journal*, *28*(4), 543-560. https://doi.org/10.1177/0265532211406422

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

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

McNamara, T. (1996). *Measuring second language performance*. Harlow, Essex: Pearson Education.

McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press.

Milanovic, M., Saville, N. & Shen, S. (1996). A study of the decision-making behavior of composition markers. In Milanovic, M. and Saville, N., editors, *Performance testing, cognition and assessment*. Selected Papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem. Cambridge: Cambridge University Press and University of Cambridge Local Examinations Syndicate, 92±114.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part 1. *Journal of Applied Measurement*, *4*(4), 386–422.

Nakamura, Y. (2004, May). A comparison of holistic and analytic scoring methods in the assessment of writing. In *3rd annual JALT Pan-SIG Conference* (pp. 45-52).

Ostovar, F. (2011). *An exploratory study on decision-making behaviors of Iranian EFL raters while holistically assessing writing tasks*. (Unpublished MA thesis). Al-Zahra University, Tehran, Iran.

Polat, M. (2020). A Rasch analysis of rater behavior in speaking assessment. *International Online Journal of Education and Teaching*, *7*(3),1126-1141.

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 15,* 18–39. https://doi.org/10.1016/j.asw.2010.01.003

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

Sakyi, A. A. (2000). Validation of scoring for ESL writing assessment: A study of how raters evaluate ESL compositions on a holistic scale. In Kannun, S. (Ed.), Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida (pp. 129-152). Cambridge: Cambridge University Press.

Sakyi, A. A. (2003). *A study of the holistic scoring behaviors of experienced and novice ESL instructors* (pp. 1230-1230). National Library of Canada= Bibliothèque nationale du Canada, Ottawa.

Shaw, S. D., & Falvey, P. (2008). The IELTS writing assessment revision project: Towards a revised rating scale (Cambridge ESOL Web-Based Research Report No. 1. Monograph). CUP/Cambridge ESOL.

Shin, S. Y., Lee, S., & Park, Y. (2023). Exploring rater behaviors on handwritten and typed reading-to-write essays using FACETS. Sadeghi, K. & Douglas, D., (Eds). *Fundamental considerations in technology mediated language assessment.* Routledge

Smith, L. J. (2008). Grading written projects: What approaches do students find most helpful? *Journal of Education for Business*, *83*(6), 325–330. https://doi.org/10.3200/JOEB.83.6.325-330

Wind, S. (2019). Do raters use rating scale categories consistently across analytic rubric domains in writing assessment? *Assessing Writing, 43,* 100416. https://doi.org/10.1016/j.asw.2019.100416.

Vaughan, C. (1991). Holistic assessment: what goes on in the rater's mind? In Hamp-Lyons, L., editor, *assessing second language writing in academic contexts*. Norwood, NJ: Ablex, 111±25.

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

Weir, C. J. (2005). Language testing and validation: An evidence-based approach. *Research and Practice in Applied Linguistics*, *9*(1), 290–301. https:// doi. org/ 10. 1057/ 97802 30514 577.

Wiseman, C. (2012). A comparison of the performance of analytic vs. holistic scoring rubrics to assess L2 writing. *International Journal of Language Testing*, *2*(1), 59-92.

Wolfe, E. W., Song, T., & Jiao, H. (2016). Features of di*ffi*cult-to-score essays. *Assessing writing***,** *27***,** 1–10. https://doi.org/10.1016/j.asw.2015.06.002

Appendices

Appendix I

General codes derived from the TAPs for all the raters using holistic rating scale

Codes	
1.	Reading the text first to get an overall impression
2.	Initial assessment of the text
3.	Anchoring
4.	Rereading the text for verifying the first impression
5.	Consulting the scale to validating his scoring
6.	Detailed analyses of the scale
7.	Revising behavior
8.	Providing evidence based on the scale
9.	Mention using of holistic scale
10.	Reading the text once adopting both scales.
11.	Praising the use of scale
12.	Reading the band scores in sequence to reach the appropriate one
13.	Assigning score based on detailed reading of the scale.
14.	Compering the scores of two rating scales
15.	Thinking
16.	Having a quick look at the scale
17.	Rereading the essay.
18.	Reading the description of band scores down up by detail and analyzing the text according to them
19.	Translating the band score description into L1
20.	Assigning score
21.	Verify his score
22.	Reading the band scores set higher than the assigned score
23.	Adopting a holistic rating scale using analytic rating scale criteria.
24.	Reading the criteria of the given score.
25.	Revise the score after reading the criteria.
26.	Providing evidence for changing the score.
27.	Looking for thesis statement
28.	Referring to scale
29.	Not reading high level band scores
30.	Reading the band score criteria to reach the appropriate level (not all of them).

Appendix II

General codes derived from the TAPs for all the raters using the analytic rating scale

Codes	
1.	First read to get the initial impression
2.	Rating the text using the scale
3.	Assigning score
4.	Silent reading of the components
5.	Validating the score
6.	Looking at the text to get a general impression.
7.	Reading the scale precisely.
8.	Looking for thesis statement.
9.	Reading the whole text.
10.	reading components' criteria down up (Not reading all the criteria completely), (not reading the excellent level)
11.	Not noticing a component before referring to scale
12.	Returning to text to look for evidence (just for specific component)
13.	Writing down the grade of each component
14.	Summing up the scores
15.	Reading the text once considering both analytic and holistic assessment
16.	Mention using analytic scale
17.	Reading the components
18.	Checking each component in the text
19.	Reading components' criteria to prove her scoring
20.	Anchoring
21.	Reading the writing sentence by sentence and make comments. (Assessing the text)
22.	Referring to the scale just for assigning score.
23.	Returning to text to read the topic after referring to scale.
24.	Assigning score to the components without reading the criteria.
25.	Having a quick look at the scale.
26.	Assigning score without considering the scale.
27.	Scoring each component using the scale.
28.	Revising his scoring
29.	Reading the whole essay several times

30.	Reading the scales several times carefully before starting the rating
31.	Prediction
32.	A general comment on the text
33.	Stating he is preferring analytic to holistic
34.	Reading the criteria in turn to reach the appropriate score
35.	Reading the rest of criteria after assigning score to the criteria
36.	Providing evidence for each components' score
37.	Stating the cause of not assigning a low score to the component
38.	Translating every description of criteria into L1
39.	Compering the scores adopting two scales.
40.	Having a quick look at the scale.
41.	Assigning score without considering the scale.
42.	Scoring each component using the scale.
43.	Revising his scoring