OPEN ACCESS

Language Education & Assessment
ISSN 2209-0959
https://www.castledown.com/journals/lea/

Language Education & Assessment, 5(1), 34–51 (2022)
https://doi.org/10.29140/lea.v5n1.777

The Processes of Rating L2 Speaking Performance Using an Analytic Rating Scale – A Qualitative Exploration

THUY THAI[a]

SUSAN SHEEHAN[a]

[a] *University of Huddersfield, UK*
lamthuy.ulis@gmail.com; s.sheehan@hud.ac.uk

Abstract

In language performance tests, raters are important as their scoring decisions determine which aspects of performance the scores represent; however, raters are considered as one of the potential sources contributing to unwanted variability in scores (Davis, 2012). Although a great number of studies have been conducted to unpack how rater variability impacts on rating decisions, much still remains unclear about what raters actually do when they assess speaking performances. This paper aimed to extend our understanding of the rating process by analysing think-aloud protocols provided by 13 VSTEP speaking raters while they used an analytic rating scale to assess 15 recorded performances with varying proficiency levels. The data suggested that although all the raters seemed to experience similar stages in the rating process, differences in rating behaviour between novice raters and experienced raters exist. To finalise the scores, the raters used five different strategies of deciding the scores.

Keywords: rater cognition; qualitative methods; high-stakes tests; analytic rating scales

Introduction

The study reported in this article investigated how raters use an analytic scale to rate speaking performances in a high-stakes test. The use of analytic scales as rating scales has gained more popularity in the language testing because analytic rating scales provide richer information about candidates' language ability since they offer more precise descriptors than holistic rating scales (Knoch, 2009). Moreover, since raters are required to pay equal attention to separate criteria (Barkaoui, 2010; Li &

Data Availability Statement: *All relevant data are within this paper.*

He, 2015), the ratings appear to be more accurate (Brown & Bailey, 1984) and can reflect more consistently the current multi-componential definition of language ability (Bachman *et al.*, 1995). However, analytic scales can be text-dense and potentially overwhelming for raters (Ballard, 2017, p. 97). In addition, raters find it difficult to have a precise understanding of the dimensions; consequently, the high cognitive load (which results from the complexity of the scale) demanded on raters can potentially lead to inconsistent ratings between and within raters (Douglas & Smith, 1997). Thus, it is of high significance to understand how raters approach the rating scales and make sense of it in their rating process. Rating process, as discussed below, is an under-researched topic.

Rating oral language is important because oral tests scores are often used for real-life, high-stakes purposes such as for employment, pay raises, placement into academic programs, and professional licensures; VSTEP speaking test scores are an example of these tests. Speaking raters also face unique challenges as they are required to quickly award a score for a test taker (TT) after the TT finish their performance and before another TT comes in. Thus, the people who rate oral assessments need to be well trained and free from personal biases while rating (Winke, 2012). It has been pointed out that "if we do not know what raters are doing...then we do not know what their ratings mean" (Connor-Linton, 1995, p. 763). Moreover, Bejar (2012) argued that rater cognition is invaluable in suggesting how best to train raters and what to monitor during the scoring process and in documenting that the mental processes raters use in assigning scores are consistent with the construct under measurement. This study covers rater cognition and documents the mental processes of the raters.

Rater behaviour and variability have gained great attention in L2 performance assessment and have been examined to provide validity evidence of test scores. A great number of studies have been conducted to unpack how rater variability impact the rating decisions, including raters' interaction with rating scales (Baker, 2012; Ballard, 2017; Brown, 2000, 2006; Winke & Lim, 2015), raters' teaching experience (Davison, 2004; Goh & Ang-Aw, 2018), raters' perceptions of fluency (Bosker *et al.*, 2014; Mulder & Hulstijn, 2011; Préfontaine, 2013), raters' perceptions of intelligibility (Deterding, 2010; Field, 2005; Levis, 2006), and raters' linguistic background (Kim, 2009; Xi & Mollaun, 2014). Nevertheless, much still remains unclear about what raters actually do when they assess speaking responses. The focus of this paper is on second language (L2) oral rating process which is an under-researched topic.

In the rating process, raters interact with three texts: the prompt, the essay/speech, and the rating scale. This paper investigates the interaction between the raters and the rating scale as the rating scale plays an important role in L2 assessment because scale content is closely related to the test construct (Fulcher, 2003). Thus, specifying what raters should attend to is important since it ultimately influences the validity of score interpretation and the fairness of decisions that educators and other stakeholders make about students based on the resulting scores (Weigle, 2002). However, Barkaoui (2010) argues that little is known about how rating scale variation affects raters and rating processes, and that such information will help improve the quality of rating scales and rater training and the test validation. This paper seeks to address this gap in the knowledge about rating scale variation.

Rater cognition and rating process have been qualitatively examined in only a limited number of isolated and exploratory studies in L2 speaking assessment, as revealed in the systematic review by Han (2016). While some studies have provided evidence that the raters differed in the way they applied and interpreted the contents of the rating scale, the others revealed contrasted findings. For example, the verbal reports of 32 official FCE examiners in Orr's (2002) study revealed that raters heeded many aspects of the performance which were not relevant to the assessment criteria; for instance, some raters' comments referred to candidate's age, gender, and candidate's presentation of her/himself. These comments seemed to influence the scores the raters decided. Although May (2006) conducted her research in the context of paired test taker interaction, the findings of the study were relevant to

this research project as it revealed the features raters paid attention to in their rating processes. May (2006) found that raters had a tendency for 'fleshing out' the criteria in the rating scale with features that were not explicitly mentioned. These aspects included features such as the first impression of the test taker, the confidence of the test taker as well as the complexity and logic of the test taker's ideas. In contrast, the findings of A. Brown's (2006) study showed almost all of the raters' comments were relevant to the scales although the raters reported some overlap between scales and some difficulties differentiating the levels. However, the issues of how raters approach analytic rating scales or descriptors and test taker responses and how they connect these two to make their score decisions were not resolved in these studies and so are addressed in this study. The identification of such features in these studies can lead to revisions of the criteria and/or rating scale and help test developers with the validation of their tests. (Knoch *et al.,* 2021) have made an important point, with which we agree, that "these sorts of behaviours can only be uncovered through the use of qualitative methods; quantitative studies fall short in this area" (p. 56).

Moreover, A. Brown and McNamara (2004) pointed out that the conflicting results in studies investigating rater behaviour and variability were perhaps not surprising, especially given that such studies tend to be small scale and exploratory, looking at a single factor at a time. They also argued that analysing the impact of specific variables such as teaching experience in isolation without considering the possible impact of other potential social identity variables, is a weakness of such studies. This point was further emphasised by Kim (2015) who concluded that rater characteristics should be carefully considered collectively in understanding rating behaviour and in training raters. One of the findings of Kim's (2015) study which was important and related to the current study was that the imbalanced attention given to the features of the descriptors was found in all three groups of raters with different background variables in the first two ratings, but this rating behaviour disappeared in the experienced raters in the third rating after they received feedback from the first two ratings. Another key finding was that experienced raters seemed to be more stable in interpreting and applying the rating scale over the three ratings than the novice and developing raters. However, novice raters in his study were defined as those who had no previous experience in the TESOL field. Thus, the difference between the two groups could be self-anticipated. Substantial differences between experienced and inexperienced raters were reported in the literature (Cumming, 1990; Barkaoui, 2010; Isaacs & Thomson, 2013; Esfandiari & Noor, 2018; etc.). For example, novice raters gave more emphasis to argumentation while expert raters put more emphasis on accuracy (Barkaoui, 2010). Similarly, novice raters were reported to have a different conception of language proficiency (Isaacs & Thomson, 2013), refer to the rating scales and rely on the criteria listed in the scales more frequently when making their scoring decisions (Esfandiari & Noor, 2018), differ more in their behaviours while assessing scripts of distinct qualities than did the medium- and high-experienced groups (Şahan & Razı, 2020) and may be more strongly affected by a particular set of criteria (Barkaoui, 2011). In contrast, raters with more experience are found to score faster (Sakyi, 2003), heed their attention to a wider variety of language features (Cumming, 1990; Sakyi, 2003), and tend to be more cautious by collecting more information before arriving at their judgments (Barkaoui, 2010; Wolfe, 1997). Additionally, Sahan and Razi (2020) suggest that raters' scoring behaviours might evolve with practice, resulting in less variation in their decisions. However, due to differences in defining expertise in the studies reviewed in the literature, mixed results are unavoidable. Therefore, the issues of what scoring behaviours are related to experienced raters and what scoring behaviours are associated with novice and developing raters remain obscure. Insights into this area are of great benefit as they can inform rater training programmes and rater certifying processes.

Furthermore, there are no cognitive processing models to reveal the underlying processes and strategies while oral raters are "attempting to understand response input, formulate a mental representation of the response, compare the response representation with that in the rating scales, and evaluate the response

in those terms" (Purpura, 2013, p. 18). These fundamental issues in rating need to be examined qualitatively to better explain raters' decision-making processes and to provide sound validity arguments for the ratings assigned by raters (Bejar, 2012; Crisp, 2012; Eckes, 2012; Wolfe & McVay, 2012). In L2 writing assessment, Lumley's (2005) study was important as it provided a detailed description of and a model of the rating process that raters followed (Table 1). The raters in his study appeared to hold similar interpretations of the scale categories and descriptors, but the relationship between scale contents and text quality remains obscure. Although rating speaking and rating writing share several features, the differences are evident. While writing raters can read texts multiple times and have more time to allocate initial scores and to consider and reconsider these scores, speaking raters cannot. Speaking raters only have one opportunity to listen to the speech while allocating and finalising their scores. Thus, we argue that the stages that speaking raters experience in their rating, which is the focus of this study, may be different from those which are proposed in Lumley's (2005) study for writing raters.

Looking at different ways of rater decision-making behaviours in L2 assessment, Cumming, Kantor, and Powers (2002) and Baker (2012) investigated the cognitive process that raters have while evaluating learners' performance. Cumming et al. (2002) made a step forward in the literature by providing a descriptive framework of the strategies employed by writing raters in their ratings. The framework suggested that raters tend to use interpretation and judgement strategies with three main focal areas, including self-monitoring focus, rhetorical and ideational focus and language focus (Cumming *et al.*, 2002, p. 88). The framework is important as it may suggest content and practice in rater training programmes and creating a clearer picture for raters to develop themselves. Baker (2012) attempted to classify writing examiners' basing on the concept of decision-making style. The analysis of 6 raters' write-aloud protocols revealed that certain elements of the texts themselves can influence raters' scoring decision. Baker also suggested that different raters may engage differently in the scoring strategies listed in Cumming et al.'s (2002) framework. More research is needed to provide more comprehensive frameworks of raters' scoring behaviours; however, research in writing assessment has moved one step further forward than research in speaking assessment.

This article seeks to address these above-mentioned issues by qualitatively investigating the rating process experienced by raters with different characteristics (rating experience and teaching experience) while rating speaking performance using an analytic rating scale in a high-stakes test of English proficiency in Vietnam.

Table 1 *Model of the Stages in the Rating Sequence (Lumley, 2005, p. 184)*

Stage	Rater's Focus	Observable Behaviours
1. 1st Reading (Prescoring)	• Overall impression of text: global and local features	• Identify script • Read text • Comment on salient features • state task requirement • give initial score impression • start to give scores
2. Rate 4 scoring categories in turn: a) TFA b) CoP c) C&O d) GC	• Scale AND text • quality of text: global and local features • features of rating scale • additional features of rating context	• Articulate and justify scores: • Nominate scoring category Refer to scale descriptors Reread text Exemplify text features Give score interruptions / recursions
3. Consider / revise / confirm scores	• Scale AND text • overall impression of text	• confirm existing scores • revise initial scores

Aim

This study was part of a bigger project which aimed at extending an understanding of factors influencing raters' scoring decisions. The intention of the current study was to examine the stages of the rating process and the strategies that the raters employed to decide the scores. It examined in detail the extent to which two groups of raters from different backgrounds used the analytic rating scale and the sequence of steps the raters followed while rating speaking performance. Specifically, the study tried to address the following questions:

Are there differences among the raters, grouped by rating experience, in

- attention paid to aspects of speaking performance
- rating stages
- score decision-making strategies?

Methods

Context of the Study

The study was conducted in the context of a high-stakes test in Vietnam—VSTEP. VSTEP, which is a test of general English proficiency, is based on the Common European Framework of Reference (CEFR) and includes a multi-level test targeting levels from B1 to C1. There are four sections assessing reading, writing, speaking, and listening, with all four sections taken by all test takers (TTs). This study focuses on the speaking component of the test.

Participants

Based on my work experience I identified raters who met the criteria I set out above and contacted seventeen raters, thirteen of whom agreed to participate and two of whom did not because of a clash of schedule, while one did not reply. The raters were teachers of English as a foreign language (EFL), who successfully attended the VSTEP speaking rater training programme and performed their rating in live test administrations for at least one year. The participants' names were anonymised with pseudonyms. The gender of the raters was not considered to be important in this study since it did not affect the rating process investigated; thus, the raters were referred as 'she' or 'her' due to the outnumber of female raters. The raters were divided into two groups according to their rating experience: experienced raters who had rated every test administration over two years and novice raters who had occasionally rated in one year. This classification criterion distinguished this study from other previous studies in that we defined rating experience as the number of ratings made by raters while other researchers defined rating experience as the years raters have worked as EFL/ESL teachers.

Instruments

Speaking test responses

With permission obtained from the relevant authorities, we gained access to 15 TTs' varying proficiency-level responses from several past VSTEP test administrations. These responses were used as rating prompts in the Think Aloud Protocol. The responses contained scripts of VSTEP speaking tests, TTs' scores and recordings. The reason for selecting these responses was to replicate the reality in which the raters worked.

Table 2 *The Participants' Information*

	Raters	Rating experience	Number of test administrations	EFL teaching experience	Students' levels
1	E1	Experienced	Over 12	over 10 years	A1, A2, B1
2	E2	Experienced	Over 12	over 10 years	A1, A2, B1
3	E3	Experienced	Over 12	over 10 years	A1, A2, B1
4	E4	Experienced	Over 12	over 10 years	A1, A2, B1
5	N1	Novice	5	over 10 years	A1, A2, B1
6	N2	Novice	5	5–10 years	B1, B2, C1
7	N3	Novice	5	5–10 years	B1, B2, C1
8	N4	Novice	6	over 10 years	B1, B2, C1
9	N5	Novice	5	under 5 years	B1, B2, C1
10	N6	Novice	4	under 5 years	B1, B2, C1
11	N7	Novice	4	under 5 years	B1, B2, C1
12	N8	Novice	5	5–10 years	B1, B2, C1
13	N9	Novice	5	5–10 years	A2, B1, B2, C1

Analytic rating scale

TT responses were scored using the 5 criteria listed in the ten-point rating scale. The criteria include Grammar, Vocabulary, Pronunciation, Fluency and Discourse management. The scale aimed to examine TTs' proficiency levels at B1, B2 and C1 according to CEFR levels.

Data Collection

The study used Think-aloud protocols (TAPs) as the only data collection method. This is a long-established method in psychological and social research and has been widely used in rater cognition literature (used in Cumming *et al.*, 2002; A. Brown, 2006, etc.). Gilhooly and Green (1996) also favoured TAPs because of its straightforwardness, without either elaboration or explanation. They continued to conclude that "such direct concurrent reports are generally accurate and reasonably complete, and have little reactive effect beyond some slowing of performance" (1996, p. 54). Data from TAPs seems to be able to best reveal the nature of raters' thinking processes when raters must talk out loud during their ratings and it relies less on memory, which does not cause cognitive strain on raters' minds and does not change the sequence of raters' thoughts (Ericsson & Simon, 1993). However, one of the most common criticisms of this method is that it tends to provide unnatural evidence on rating behaviour (Ballard, 2017). In other words, the verbalisation may change the nature of the process (Stratman & Hamp-Lyons, 1994). For example, raters may focus more equally on all criteria in the rating scales when being asked to talk through their rating processes while they may not do similarly in their regular practice. In order to solve this, TAPs in our study were conducted on two occasions in which the 13 raters rated 15 speaking performances all together. This seems to be the largest number of speaking performances rated by the largest number of the raters using TAPs in the L2 assessment literature to date. Due to time constraints and cognitive load to perform the task, the raters rated the responses in two occasions. The raters rated 6 responses in the first occasion and rated the other 9 responses in the second occasion. The time gap between the two occasions was 1–2 weeks so that the participants did not forget the TAPs procedure and that they would get familiar with the TAPs, thus being closer to the nature of the process.

Table 3 *Summary of Data Volume*

	Raters	Duration of TAPs assessing 15 performances (hour:minute: second)	Word count
1	E1	4:05:41	15,869
2	E2	2:34:05	10,438
3	E3	4:40:00	16,667
4	E4	4:18:58	13,271
5	N1	3:51:53	4,850
6	N2	4:14:32	16,698
7	N3	3:39:16	12,803
8	N4	3:05:30	6,641
9	N5	3:27:39	6,820
10	N6	3:53:34	7,201
11	N7	2:51:14	4,156
12	N8	3:13:48	4,061
13	N9	2:37:18	6,840

Prior to the rating session, the 13 raters went through a training session of TAPs since the training allowed the participants to be familiar with the required procedure and clarify any misunderstandings. The training session adopted the procedure suggested by Ericsson and Simon (1993, pp. 16–18). The participants, then, were arranged to sit in a quiet room in an authorised area in order to avoid unnecessary interruption by "other voices or external disturbances" (Richardson, 1996, p. 59). Table 3 summarises the duration of each rater's recorded TAPs and the number of words when it was transcribed.

Data Analysis

All the TAPs recordings were transcribed and segmented into meaningful units. Segmentation of rater talk posed a problem with no ideal answers. After a number of considerations, it was decided to adapt Lumley's (2005) intuitive approach, in which initial segmentation was guided by a small number of principles to allow analysis of:

- the rating sequence, later divided into three main stages: initial attention, score allocation and score decision
- the features considered under each assessment criterion, including Grammar, Vocabulary, Pronunciation, Fluency and Discourse management
- the strategies the raters used when making final decisions
- other features of raters' talk in their TAPs

Finally, a coding scheme was fully developed as the data analysis went on. The TAPs of each rater were coded into the following themes.

Findings

All the raters appeared to encounter three main stages in their ratings. Once the TTs started talking, the raters started paying attention to aspects of the speaking performance specified in the rating scale such as accuracy of grammar, sophistication of vocabulary, etc. Next, the raters allocated initial scores for

Initial attention paid to each assessment criterion	(1) Grammar	(1) Accuracy (errors) (2) Complexity
	(2) Vocabulary	(1) Accuracy (word choice, word form) (2) Range (3) Lexical sophistication
	(3) Pronunciation	(1) Individual sounds (2) Stress (word level & sentence level) (3) Intonation
	(4) Fluency	(1) Hesitation & Pauses (2) Speed (3) Error correction
	(5) Discourse management	(1) Cohesive devices (2) Relevance of ideas (3) Idea elaboration
	(6) Others	(1) Test questions (2) Interlocutor manner (3) TT age
Score allocation	(1) Grammar (2) Vocabulary (3) Pronunciation (4) Fluency (5) Discourse management (6) Others	
Score decision	(1) Grammar (2) Vocabulary (3) Pronunciation (4) Fluency (5) Discourse management (6) Overall performance (7) Others	

one or several assessment criteria (e.g., Grammar and/or Pronunciation). Then they made their final decisions of scores at the end of the TTs' talks. These stages were discussed in detail below.

Attention Paid to Aspects of Speaking Performance

It was evident in the data that the rating scale exerted an impact on what the participant raters attended to in their ratings perhaps as a result of the rater training programmes they received. Almost all of the comments were about the criteria including Grammar (Gr), Vocabulary (Vo), Pronunciation (Pr), Fluency (Fl) and Discourse management (DM), none were neglected. Very few comments were identified as "others" when they were not related to the rating scale. These comments occasionally referred to the difficulty of the test questions, the interlocutors' manners and the TT's age. A useful illustration of how the raters paid attention to the speaking performance is provided in Excerpt 1, which is drawn from E1's TAPs. Her first comment was on the complexity of the grammatical structures that the TT used at the very beginning of his/her talk. The following comments focused on different aspects of the speech rather than on one aspect.

(1)

E1 – TT2

1 The TT can use complex structures	Gr – Complexity
2 and extend the ideas for the first question	DM – Thematic development

3	The TT can use the structure with "because", "and", "when"	Gr – Complexity
4	and there is one grammar mistake related to the use of "very" before main verb.	Gr – Accuracy
5	A bit concerned that the interlocutor asked a question not mentioned in the script.	Interviewing behaviour
6	The TT can use the word "balance"	Vo – Sophistication
7	although making a mistake—using "many" with an uncountable noun	Vo – Accuracy

The five assessment criteria received unequal attention from the raters as the data shows. For example, E1 paid more attention to Grammar than to any other of the remaining 4 criteria. This rating behaviour of paying unequal attention to the assessment criteria occurred with the other raters as well. One possible explanation for this was that some aspects of the speaking responses were more salient to some raters than others. This, perhaps, was due to the fact that different raters had different perceptions of the criteria in assessing speaking. E1 may have prioritised Grammar over other criteria. It deserves a note here that E1 seemed to spend most of her teaching time with low-proficiency and non-English major students. Their entry levels to her class were A1 to A2 level, according to CEFR levels and their exit level were set to be A2 or B1, equivalently. With these levels, the students are known to be developing their English, thus making grammar mistakes was unavoidable. Therefore, her EFL teaching experience with certain groups of students may have allowed her more sensitivity to comment on the TTs' use of grammatical structures.

All the raters paid attention to all the descriptors in the rating scale; however, their attention was not paid equally to those descriptors. For example, in assessing Pronunciation, the raters paid attention to all the descriptors in the rating scale, including individual sounds, word stress, sentences stress, and intonation; however, they all seemed to overwhelmingly attend to the ending sounds of the words the TTs pronounced (see Excerpt 2).

(2)

N7 – TT4

1 Lack of ending sounds
2 Club (should be clubs)
3 many time (should be times)
4 this TT can't pronounce the ending sound—language
5 this TT pronounced "dance"—fail to pronounce the ending sound

This type of comments was typical among all the raters. The final sounds of words received much more attention from the raters than other features of pronunciation. It might be due to the fact that Vietnamese speakers do not have to pronounce the ending sounds of words; thus, naturally Vietnamese learners of English tend to delete or substitute most of them. Consequently, this common feature might have been more identifiable to the raters in the study. This appear to indicate that the attention the raters pay toward the speech can be conditioned by the context they are in and the teaching experience they have.

The data seems to suggest that the raters with more rating experience paid more sophisticated evaluative attention to the speech features than those with less rating experience. Two points are noteworthy here. First, the experienced raters provided more comments and that their comments were more specific. In contrast, novice raters tended to provide general evaluative comments.

Second, the process of assessing language features which are represented by the descriptors in higher band scores seemed to be easier for experienced raters than the novices. One example taken from among the data which exemplifies these rating features is the way that the raters commented on "*less common words and idiomatic expressions*" (language in the descriptors) in the TTs' speech (Excerpt 3).

(3)

E1 – TT3

1 Oh this TT can use a very good phrase—part and parcel
2 She has a good way of using the vocabulary item, I see eye to eye with
3 However, I have the feeling that she learns by heart some structures to get higher band scores, but her ability to elaborate the ideas is not shown in this answer.
4 Because it is clear that I expect her to explain what she "sees eye to eye with", but the later part of her answer did not have any relevance (smile) to the idea she stated earlier.
5 This TT gives me a feeling that she has practised the test-taking strategies. It means she knows how to include some structures to make the raters think that she can be at a very high level.
6 But the way that she elaborates the idea or the grammar mistakes and word forms and word choices she made reveal that she cannot reach the level that she can actually use those structures.

N8 – TT3

1 part and parcel, is this an idiom?
2 see eye to eye?

Excerpt 3 illustrates that the similar rating feature among all the raters was that they often repeated the words/phrases with pleasant voice when they heard some "highlights" in the use of vocabulary. However, E1, an experienced rater with more than 10 years of teaching experience provided a detailed account of the TT's use of the idiomatic expressions. E1 also seemed to be confident in identifying the inauthentic use of the expressions due to "*the test-taking strategies employed by the TT*" while N8, a novice rater with less rating experience and less than 10 years of teaching EFL, appeared unconfident about both identifying the idiomatic expressions and evaluating the appropriateness/authenticity of their use. This difference in the raters' rating behaviour occurred frequently in the data. This finding seems to indicate that there might be some degree which rating and teaching experience can play a role in enhancing rating confidence in raters. Thus, it is important to consider rater characteristics in extending understanding of rater behaviour in rater-focused research.

Allocation of First Scores

The second stage that the raters experienced in their ratings was allocation of the first scores. The first scores in this section mean the score(s) that the raters nominated initially in their TAPs for each TT. Generally, all the raters except E1, E3, N4 and N9 seemed to allocate their first scores either in or after the TTs finished answering part 1 questions. There was no particular order of the criteria for the allocation nor time of the allocation; sometimes they allocated single score for one criterion first, in other times they allocated several criteria at the same time, and in some other times they allocated the overall score for the performance. This feature is illustrated in Excerpt 4.

(4)

Allocation of single score

1 computer (wrong pronunciation)
2 ending sound
3 pronunciation probably band 4 because…easy to listen to, clear pronunciation, but individual sounds particularly ending sounds not correct (TT7)

Allocation of several criteria score

1 Grammar…at least band 6. Enough to answer part 1 questions clearly and easily, express quite many ideas.
2 Pronunciation is quite easy to listen to, natural, not yet seen any prominent issues, so at least band 6…band 5 or band 6 for pronunciation.
3 Fluency is quite ok, the speed to respond to the questions is quite fast and keep talking. Hesitation is not obvious, stop at band 6 at least.
4 Discourse marker, oh discourse management, after part 1, express ideas with quite a lot of supporting details, attempt to elaborate on ideas and organise ideas, particularly the question about advice, so at least band 6. (TT11)

Allocation of overall score

Only single-word answers. Definitely can't have band 5. (TT5)

These typical examples of the ratings may indicate three features of the raters' behaviour. First, they did not give separate attention to each of the rating criteria when they nominated a rating criterion and allocated the score. This may have been due to the fact that the nature of the speech in speaking performance and the rating time tension may not allow speaking raters to nominate each criterion at one time. Second, the raters seemed to allocate the scores quickly, with relatively little deliberation, and expressed little or no uncertainty about the scores they nominated. This is typical of what occurred in the data. They usually added some hints of why they came up with a certain number although they seemed not to have much explicit reference to the rating scales. By this we mean, they did not always use the words in the rating scale in explaining their score allocation. Perhaps they had internalised the rating scales and/or they may have held a particular expectation of what a band may look like. Third, the raters seemed to occasionally use their holistic evaluation as a reference when nominating their first scores instead of referring to individual assessment criteria presented in the rating scale. The holistic evaluation appeared to help the raters narrow down the range of band scores that they were looking at during their ratings. This suggests that the role of the rating scale was as a classificatory scheme for the raters' impression of the speech.

It seems that the raters with less rating experience found themselves puzzled in allocating scores more often than those with more rating experience. For example, there were several situations in which novice raters nominated several scores for one criterion when they said: "I was quite concerned, considering if it's a 3,4 or 5".

Score Decision-Making Strategies

There seemed to be five strategies which were employed by the raters while they made their score decisions. However, the raters with more rating experience seemed to judge their final scores for all the assessment criteria in the order they were presented in the rating scale while it was less frequent to see those with less rating experience provide verbal justification of the scores they decided. Additionally,

raters with more rating experience seemed to be more confident in making their scoring decisions while novice raters occasionally found it difficult to decide the scores.

Matching

One of the most common strategies that the raters used was matching the speech features with the descriptors in the rating scale to justify their score decisions. The examples below illustrate two typical ways of experienced raters and novice raters when they decided their scores (Excerpt 5)

(5)

E1 – TT14
Vocabulary is really her strong point. She can use good command of broad vocabulary. She can use less common words, right? But she still makes mistakes in word choice and word form, so she can't have a 9…because it's not minor slip, right? If it's a 9, it must have no significant lexical error, so I give this TT a 8 for vocabulary.

N6 – TT1
Grammar, band 3, use simple structures, systematic basic mistakes, ok but can be understood. But this TT has one thing, that is the attempt to use complex sentences, so 3 and 5, in total, Grammar has many mistakes but I have to give it a 4.

The excerpt illustrates the way both raters were trying to convince themselves of the best possible score for the TT. Both E1 and N6 explicitly referred to the descriptors by reading out loud the key words/phrases. Beside mentioning both positive and negative points, the matching strategy also involved the rejection process of other possible bands. This seems to indicate the significant role of the rating scale in the raters' final decision.

Moreover, in their final judgement of the scores, the raters seemed to consider both positive and negative factors although N6 tended to pay more attention to the grammatical accuracy than aspects of grammar described in the rating scale. Additionally, it can be seen from the examples that E1 was more specific than N6 in justifying her decision by pointing out the complex structures that the TT could use. This type of difference between experienced and inexperienced raters appeared to occur frequently in the data.

Regarding the times when the raters gave their justification for the final scores, the raters with more rating experience, including E1, E3, E4, E2 always verbalised their score decisions for all the assessment criteria as they were sequenced in the rating scale after the TTs finished their performances. In contrast, those with less rating experience tended to decide the scores for several assessment criteria during the TTs' talk and the judgements did not follow the sequence in the rating scale. For example, N3 started her score decisions with Fluency, the fourth criterion in the rating scale, while the TT was performing toward the end of his/her performance. She said: *"Fluency not good…just a bit over band 3. Band 3 means "noticeable hesitation, frequent false starts, and repetition. So band 4 seems not confident. Time's up. Not much was talked."* This may indicate that the sequence of the rater's judgment was sometimes conditioned by the rater's perception of the salience of one or two aspects of the performance. It would have been easier for those with less rating experience to decide the scores for those salient features, then they would move to other less salient features.

Furthermore, there were some occasions in which the novice raters found themselves torn between two band scores and it took them a long time to decide the scores. For instance, two long pauses were

found when N8 had to make the decision of scoring either a 5 or a 6 for Discourse management for TT8. Although she referred to the descriptors of both band 5 and 6, she seemed unable to convince herself of the best score for the TT by explaining that "*in the last part when she [the TT] talks about better…something like widen knowledge of culture, she does elaborate more on. A 6 is not a full 6… and the linking words she can have some complex connectors, lexical linking words like 'another advantage is'…[sigh]*." It can be seen that she herself was not satisfied with her own justification, but it might have been due to time limit she finally decided a 6. This situation occasionally occurred in the data of N5 and N9, who were also novice in their ratings, thus providing further evidence in support of the notion that raters with different rating experience differ in the way they perform their ratings and that novice raters tended to be less confident in using this strategy than experienced raters.

Simplifying key terms in the descriptors

Another strategy that the raters used while giving their final judgements on the scores was simplifying key terms in the descriptors (Excerpt 6).

(6)

N5
This TT shows quite enough vocabulary for the content she needs to talk about although she makes quite many mistakes in vocabulary and sometimes has difficulty in expressing with vocabulary or repeats many times.

N3
Pronunciation is quite easy to listen to, no obvious issue or heavy strain on listeners. This TT's problem is being unable to talk much, no vocabulary and ideas to talk…so unable to talk, but it does not mean poor pronunciation. So a 4 for pronunciation.

The raters tended to simplify the key terms in the descriptors by using their own words, then on which they based to justify and confirm their scores. In these instances, N5 used "quite enough" to possibly refer to "sufficient vocabulary" or she used "difficulty in expressing with vocabulary" to possibly mean "difficulty with unfamiliar topics and make many lexical errors" in the descriptors. Similarly, N3 used "no obvious issue or heavy strain on listeners", which may refer to "generally clearly articulate individual sounds". This seems to be one of the raters' strategies to deal with the complexity of language in the rating scale. Moreover, the language used seemed to be more informal than that of the rating scale. It was as though the raters needed to re-express the wordings of the scale in their own terms to make it fit more closely to their gut reaction to the speech. This may suggest the scale wordings did not adequately describe their attitude to the speech, but if reinterpreted in this way, they were close enough for the raters to accept and use them.

Referencing to holistic rating

Although VSTEP rating scale is an analytic one, all the raters seemed to frequently mention the overall proficiency level of the TTs (Excerpt 7).

(7)

N6: 4 4 4 3 5, average a 4, B1 is probably good enough. (TT1)
N2: but this TT is surely C1, so to balance I will give two 9s, one for fluency and the other for discourse management. (TT15)

This seems to indicate that the raters tried to reconvince themselves of their final decision by basing on the overall proficiency level that they thought the TTs deserved. Holistic scoring seemed to play an important part in the raters' rating decisions. This finding is important as it allows an insight into the process of speaking assessment, suggesting that this is an area for further research as holistic evaluation had previously been assumed unpresent in the use of analytic rating scale.

Compensating

Another decision strategy found in the data is compensating, which means the raters would try to balance the scores of two criteria or more if they found that the TTs could reach a higher band score for one criterion but a lower band score for the other criterion. For example, when E2 thought a 5 for Discourse management would be high for TT12, she lowered the score for that criterion (which was finally a 4) and lifted the score for Fluency as a 5 rather than a 4. This strategy was commonly applied by other raters throughout their TAPs. The application of this strategy appears to indicate that VSTEP raters tried their best to bring the benefits in terms of scores to the TTs. They were trying to do the best for the TTs. This nurturing aspect in their ratings was shown in the way they treated the borderline performances as well. Perhaps as a teacher, the raters seem to understand the difficulty that a learner of English had to encounter particularly in their context. They may also understand how hard it was to achieve the higher band score and the significance of the scores to the TTs.

Using own sense

The last strategy that the raters used in their scoring decisions is using their own sense, which means it is difficult for them to articulate the reasons why they arrived at that score. The following example illustrates this decision-making strategy (Excerpt 8).

(8)

E1
This TT has some complex connectors, such as besides or first of all, but I do not feel that she can have a 5, so I give her a 4 for this criterion.

Sometimes instead of explaining the reasons for the scores or using other strategies mentioned above, the raters used their own sense which was probably accumulated through their rating and teaching experience to come to the final scores. N2 made a similar judgement in her TAPs when she said: "*Although I feel it is a bit high [for the TT], during the rating time I decided a 6 for her pronunciation*." Although there is variation amongst raters in the frequency of this kind of comment, they all encountered this problem, and overall, it is a common response. What seems to take place in these examples is a struggle between an internal, intuitive sense of the value of each score level, and the public articulation of the score in terms of the scale, which may seem unsatisfactory as a description of the speech. In the end the raters in this study tended to rely on their intuitive sense for their final decision in these cases. This can indicate that the scale is sometimes inadequate to the complexity of what the raters observed, which inevitably leads to a tension between reliability, represented by the rating scale levels and the impression the raters had gained of the speech.

Discussion

This study qualitatively examined the rating process of raters with different backgrounds. The two groups of raters displayed similar rating stages: paying attention to the speech features described in the rating scale, then allocating the initial scores and deciding the final scores. They all tended to use five

decision-making strategies in their ratings: matching, simplifying key terms, referencing to holistic rating, compensating, and using their own sense. This finding is important as it contributes new knowledge to the body of literature by revealing the underlying processes and the strategies used by speaking raters. Very little was found in the literature on the question of what the mental processes of rating speaking, or the rating sequence was. Most of the rating processes and strategies are proposed through investigation in rating writing (e.g.: Cumming *et al.,* 2002; Baker, 2012; Lumley, 2005). This study has illustrated that the stages that speaking raters experience are different from the stages experienced by writing raters in Lumley's (2005) work. The VSTEP speaking raters started their rating process without a pre-scoring stage as used by writing raters. The participant raters' focus was heavily influenced by the assessment criteria listed in the rating scale, which is similar to the writing raters in Lumley's work. However, the raters in this study did not pay separate attention to the assessment criteria in turn due to the fact that they were allowed to listen to the speech only once. They attended to several criteria at one time. Moreover, some participant raters retained the speech features in their mind while rating, whereas others took notes of the evidence on paper so that they would not forget it. This rating behaviour is different from that of writing raters since the text features are displayed in the texts to which the writing raters can refer at any time during their rating. One similarity was found at the last stage when the raters confirm their scores. Both the participant raters in this study and Lumley's raters referred to their overall impression when arriving at the final scores. The similarities and differences among raters in the mental processes of rating writing and speaking performances might be indicative of a need to reconsider to what extent findings in writing rater behaviour research can be applicable to those in speaking rater behaviour. This study also sheds light on the need for further investigation into the rating processes experienced by speaking raters. This is an area of research which currently receives less attention than similar issues in writing assessment.

Additionally, the finding of the use of holistic rating by the raters was important, despite the fact that holistic rating seems not to be mentioned in discussion of rating with an analytic rating scale in the literature. The participant raters in the current study referred to their holistic rating in different ways. E1, E2, E4, N2, N3 and N6 referred to holistic rating as a confirmation of their score decisions while the others referred to it as a way of allocating their first scores. The finding should be interpreted with caution as it is evident in the data that the raters' rating behaviour was conditioned by the analytic rating scale. The holistic rating seemed to be an added reference which increased the raters' confidence in terms of scoring. If rater confidence was a desirable characteristic (as suggested by Cushing, 2019), it would be important to consider whether referencing to holistic rating is a construct relevant factor and then how this reference could be incorporated in analytic rating scales. On this point, Bejar (2012, p.6) wrote that:

> Clearly, then, rater cognition is central to a validity argument involving scores based on human scores. When such scores are, in turn, the basis for other scores or products, rater cognition remains relevant because it is the foundation, or at least a component of, their corresponding interpretive arguments.

Another key finding of this study was the differences between experienced and novice raters. First, experienced raters appeared to pay more sophisticated attention toward the speech features. This rating behaviour was similarly found in other previous studies (Cumming, 1990; Isaacs & Thompson, 2013; Sakyi, 2003). Second, the participant raters with more rating experience appeared to be more confident in using the matching strategy than those with less rating experience. Moreover, the study also has implications for rater training as it documented the mental processes the raters used in assigning scores. Raters with different backgrounds seem to display different rating behaviour in their rating process. Thus, it might be helpful to provide them with individualised feedback to help them more confident in performing their rating job, rather than one-size-fits-all training programmes for all raters.

While this study has revealed more insights into the process of rating L2 speaking performance, the study has yet little evidence of whether certain of these rating behaviours are associated with better raters. Such further research may be needed to indicate whether certain rating behaviours are more or less desirable, thus allowing them to be built into training models (Brown, 2016, p. 421). In addition, further insights into the rating processes may be revealed if the researchers could have interviewed the raters about their opinions on the rating scale, their approach to it and reflection on their rating practice. Furthermore, since the way the raters rated may have been influenced by the training sessions they have received, it may be important to examine the role of training sessions in shaping rater behaviour, together with other variables.

This study has generated new insights into the process of speech rating. These insights have significance beyond the Vietnamese context as analytic scales are commonly used in tests around the world. Speech rating was an under-researched topic in comparison with writing rating. This study has addressed this gap in the literature and demonstrated that the rating processes are different for the different language skills. Furthermore, this study has shown that the rating process is different for novice and experienced raters. Previous studies have examined individual rater characteristics such as teaching experience. This study, in contrast, has looked at raters in their myriad complexities. The potential impact on the field lies in the insights that raters need more individualised training and the differences between raters should not only be studied in terms of considering if a rater is fitting or not in relation to scores given by the majority of raters. This study has provided insights into what raters do when assessing responses and the processes used to determine the TT scores. The connections between the rating scales and the TT responses have been uncovered. In sum, this study has generated insights into the rating processes of oral raters.

References

Bachman, L., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing, 12*(2), 238–257.

Baker, B. A. (2012). Individual differences in rater decision-making style: An exploratory mixed-methods study. *Language Assessment Quarterly, 9*(3), 225–248. https://doi.org/10.1080/15434303.2011.637262

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

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

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

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice, 31*(3), 2–9.

Bosker, H. R., Quené, H., Sanders, T., & de Jong, N. H. (2014). The perception of fluency in native and nonnative speech. *Language Learning, 64*(3), 579–614. https://doi.org/10.1111/lang.12067

Brown, A. (2000). An investigation of the rating process in the IELTS Oral interview. *IELTS Research Reports, 3.* Retrieved from https://www.ielts.org/-/media/research-reports/ielts_rr_volume03_report3.ashx

Brown, A. (2006). An examination of the rating process in the revised IELTS Speaking Test. *IELTS Research Reports, 6.* Retrieved from https://www.ielts.org/-/media/research-reports/ielts_rr_volume06_report2.ashx

Brown, A. (2016). Interlocutor and rater training. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 413–425). Routledge.

Brown, A., & McNamara, T. (2004). "The devil is in the detail": Researching gender issues in language assessment. *TESOL Quarterly, 38*(3), 524–538. https://doi.org/10.2307/3588353

Brown, J. D., & Bailey, K. M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning, 34*(4), 21–38. https://doi.org/10.1111/j.1467-1770.1984.tb00350.x

Connor-Linton, J. (1995). Looking behind the curtain: What do L2 composition ratings really mean? *TESOL Quarterly, 29*(4), 762–765. https://doi.org/10.2307/3588174

Crisp, V. (2012). An investigation of rater cognition in the assessment of projects. *Educational Measurement, Issues and Practice, 31*(3), 10–20. https://doi.org/10.1111/j.1745-3992.2012.00239.x

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

Cushing, S. (2019). Speaking and writing rater training. Pre-conference workshop presented at the 6th AALA Conference, Hanoi, Vietnam.

Davis, L. (2012). *Rater expertise in a second language speaking assessment: The influence of training and experience.* Unpublished doctoral dissertation, University of Hawai'i at Manoa.

Davison, C. (2004). The contradictory culture of teacher-based assessment: ESL teacher assessment practices in Australian and Hong Kong secondary schools. *Language Testing, 21*(3), 305–334. https://doi.org/10.1191/0265532204lt286oa

Deterding, D. (2010). Norms for pronunciation in Southeast Asia. *World Englishes, 29*(3), 364–377. https://doi.org/10.1111/j.1467-971X.2010.01660.x

Douglas, D., & Smith, J. (1997). *Theoretical underpinnings of the Test of Spoken English revision project.* Educational Testing Service.

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

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

Esfandiari, R., & Noor, P. (2018). Iranian EFL raters' cognitive processes in rating IELTS speaking tasks: The effect of expertise. *Journal of Modern Research in English Language Studies, 5*(2), 41–76. https://doi.org/10.30479/jmrels.2019.9383.1248

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly, 39*(3), 399–424.

Fulcher, G. (2003). *Testing second language speaking.* Pearson Education.

Gilhooly, K., & Green, C. (1996). Protocol analysis: theoretical background. In J. T. Richardson (Ed.), *Handbook of qualitative research methods for psychology and the social sciences* (pp. 43–54). BPS books.

Goh, C. C. M., & Ang-Aw, H. T. (2018). Teacher-examiners' explicit and enacted beliefs about proficiency indicators in national oral assessments. In D. Xerri & P. V. Briffa (Eds.), *Teacher Involvement in high stakes language testing* (pp. 197–215). Springer.

Han, Q. (2016). Rater cognition in L2 speaking assessment: A review of the literature. *Studies in Applied Linguistics & TESOL, 16*(1), 1–24. https://doi.org/10.7916/salt.v16i1.1261

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly, 10*(2), 135–159. https://doi.org/10.1080/15434303.2013.769545

Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. *Language Assessment Quarterly, 12*(3), 239–239. https://doi.org/10.1080/15434303.2015.1049353

Kim, Y.-H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing, 26*(2), 187–217. https://doi.org/10.1177/0265532208101010

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

Knoch, U., Fairbairn, J., & Jin, Y. (2021). Scoring second language spoken and written performance: issues options and directions. Equinox Publishing Ltd.

Levis, J. M. (2006). Pronunciation and the Assessment of spoken language. In R. Hughes (Ed.), *Spoken English, TESOL and applied linguistics* (pp. 245–270). Palgrave Macmillan. https://doi.org/https://doi.org/10.1057/9780230584587_11

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

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

Mulder, K., & Hulstijn, J. H. (2011). Linguistic skills of adult native speakers, as a function of age and level of education. *Applied Linguistics, 32*(5), 475–494. https://doi.org/10.1093/applin/amr016

Orr, M. (2002). The FCE Speaking test: using rater reports to help interpret test scores. *System, 30*(2), 143–154. https://doi.org/10.1016/S0346-251X(02)00002-7

Préfontaine, Y. (2013). Perceptions of French fluency in second language speech production. *The Canadian Modern Language Review / La revue canadienne des langues vivantes, 69*(3), 324–348. https://doi.org/10.3138/cmlr.1748

Purpura, J. E. (2013). Cognition and language assessment. In A. J. Kunnan (Ed.), *The companion to language assessment.* Wiley https://doi.org/doi:10.1002/9781118411360.wbcla150

Richardson, J. T. E. (1996). *Handbook of qualitative research methods for psychology and the social sciences*. British Psychological Society.

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

Sakyi, A. A. (2003). *A study of the holistic scoring behaviours of experienced and novice ESL instructors.* Unpublished doctoral dissertation, University of Toronto.

Stratman, J. F., & Hamp-Lyons, L. (1994). Reactivity in concurrent think-aloud protocols: Issues for research. Speaking about writing: *Reflections on Research Methodology, 8*, 89–111.

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

Winke, P. (2012). Rating oral language. In *The encyclopedia of applied linguistics*. Wiley Online Library. https://doi.org/10.1002/9781405198431.wbeal0993.pub2

Winke, P., & Lim, H. (2015). ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study. *Assessing Writing, 25*, 37–53. https://doi.org/10.1016/j.asw.2015.05.002

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing, 4*(1), 83–106. https://doi.org/10.1016/S1075-2935(97)80006-2

Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational measurement, issues and practice, 31*(3), 31–37. https://doi.org/10.1111/j.1745-3992.2012.00241.x

Xi, X., & Mollaun, P. (2014). How do raters from India perform in scoring the TOEFL iBT speaking section and what kind of training helps? *ETS Research Report Series.* https://doi.org/10.1002/j.2333-8504.2009.tb02188.x