Factors influencing raters' scoring decision and their rating practice development: A study of a high-stakes test in Vietnam

Thuy Ha Lam Thai

A thesis submitted to the University of Huddersfield in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The School of Education and Professional Development

December 2021

Copyright statement

Acknowledgements

First, I would like to thank my supervisor, Susan Sheehan. I am deeply grateful for her generosity with her time, her critical yet constructive feedback, and her unwavering support. Her firm commitment to her students is truly inspirational. I would like to thank my second supervisor, Aylin Unaldi, who provided me with helpful feedback on the messy first draft of the whole text. It was a huge pity that I had not had her at the start of my journey.

The thesis would not have been completed without fourteen raters who participated in my research project. I am extremely grateful for their willingness to participate in this study and their extraordinary generosity with their time.

To my son and husband, my special thanks are for the encouragement and the space you provided throughout.

Abstract

This study explores the issue of rating speaking from a so far underused qualitative research perspective. This qualitative study investigated the experiences of Vietnamese-L1 raters in scoring a localised high-stakes test of English proficiency, Vietnamese standardised test of English proficiency (VSTEP). There were three overarching aims that the thesis aimed to achieve:

- To further an understanding of the rating process experienced by Vietnamese raters in the assessment of speaking performance in English as a foreign language
- To further an understanding of the factors which affect the raters' scoring decisions
- To further an understanding of the ways raters develop their rating practice over time

All of the 14 participants were Vietnamese teachers of English as a foreign language (EFL), had attended rater training programmes and had rated for at least a year. First, all of the participants were invited to participate in a moderation discussion when the participants rated three bench-marking performances and discussed why they arrived at their scores. After that they were asked to use think-aloud protocols (TAPs) to rate 15 VSTEP speaking performances under the audio rating condition. The participants were then interviewed using a semi-structured approach. Data from moderation discussions, TAPs and interviews were all recorded and then analysed using interpretative phenomenological analysis.

The findings suggested that there were individual differences in perceptions of assessment criteria, rating procedure, score decision time and score decision strategies among the raters though they were recruited from a homogenous

group. However, they all experienced a three-stage process while rating: consciously paying attention to the speech features, allocating their first scores and finalising the scores using 5 different decision-making strategies. There existed a number of differences which distinguished novice raters from experienced ones. For example, experienced raters provided more and longer comments on test-takers' speech than novice raters and with more confidence. Regarding the factors contributing to score variations, the treatment of local speech features, the English standard(s) and the perception of syntactic and lexical accuracy and complexity and discourse competence seemed to play significant roles in generating such differences. The data also showed that the rating experience and the number of trainings contributed considerably to the raters' rating development. The study reveals the three stages of how novice raters evolved into experienced raters.

The results extend our understanding of the rating process in the assessment of speaking performance in EFL and how rating behaviours evolved with training and practice. The study also contributes to a better understanding of local raters' perceptions and practices of assessing English in their local context. There were significant implications which could be drawn from this study for speaking constructs assessed in a localised EFL context. Moreover, the study has implications for the enhancement of the effectiveness of rater training programmes and the standardisation procedures.

Table of contents

Table of Contents

List of figures and tables

Tables

List of abbreviations

APTIS: an English language test developed by British Council

CET: College English Test

CEFR: Common European Framework

CoP: Community of Practice

E: Experienced (raters)

EFL/ESL: English as a foreign/second language

ELF: English as a lingua franca

EIKEN: Jitsuyo Eigo Gino Kentei (Test in Practical English Proficiency)

GCSE: General certificate of secondary education

IELTS: International English language testing system

IPA: Interpretative phenomenological analysis

L1: First language

L2: Second language

MD: moderation discussion

N: Novice (raters)

NFL2020: National foreign languages project 2020

SR: stimulated recall

TAPs: Think-aloud protocols

TEAP: Test of English for academic purpose

TOEFL: Test of English as a foreign language

TT: Test taker

VSTEP: Vietnamese standardised test of English proficiency

Chapter 1: Introduction

This thesis explores the lived experience of Vietnamese trained raters in a high-stakes English speaking language proficiency test. It is concerned particularly with mental processes while rating, the factors that influence raters' scoring decisions on test-takers' (TTs) performances and how they have developed their rating practice from their own experience of working with the test. This introduction explains how this research focus emerged, and details the overall aims of the study, concluding with an outline of the structure of the thesis. It also presents the context in which the study was conducted, and the impact on this of changes to high-stakes language testing in a local context.

1.1 Background of the study

In tests of speaking performance, it is important that the people who perform oral assessments are well trained and free from personal biases while rating since the scores assigned have important consequences for test reliability and validity (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 research into raters' mental processes could provide invaluable insights to inform rater training programmes and rater monitoring measures during the scoring process and to document whether the mental processes raters use in assigning scores are consistent with the construct under measurement.

Understanding the significance of the rating task, rater behaviour and variability have gained great attention in second language (L2) performance assessment. A great number of studies have been conducted to unpack how rater variability impacts on rating decisions (see chapter 2 for detailed discussion), including raters' interaction with rating scales (B. A. Baker, 2012; Ballard, 2017; A. Brown,

2000, 2006; Winke & Lim, 2015), raters' teaching experience (Davison, 2004; Goh & Ang-Aw, 2018), raters' perceptions of fluency (Bosker, Quené, Sanders, & de Jong, 2014; Mulder & Hulstijn, 2011; Préfontaine, 2013), raters' perceptions of intelligibility (Deterding, 2010; Field, 2005; Levis, 2006), and raters' linguistic backgrounds (Y.-H. Kim, 2009; Xi & Mollaun, 2009). Nevertheless, much still remains unclear about what raters actually do when they assess speaking performances, in what ways these factors interact to influence raters' scoring decisions and how their rating practice develops over time. The reason for this may lie in the differences in methods used, the participants recruited, and the rater-related factors investigated. For example, while A. Brown (2000, 2006) used verbal reports as the main method to examine the rating behaviour of trained raters in an international English language testing system (IELTS) test, Davis (2016) investigated the impact of training and experience on scoring decisions of experienced teachers of English by using mixed methods. In addition, the research to date has tended to focus on single factors rather than considering how these sets of factors influence raters' scoring decisions. As argued by A. Brown and McNamara (2004), 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 (more discussion in sections 2.3, 2.4 and 2.5). I would argue this is a gap that needs to be filled by empirical studies.

Therefore, adequate understanding of how raters arrive at the scores and why they behave the way they do in their rating process is of considerable benefit. First, knowledge of the rating sequence, the decision-making strategies raters use and the language features raters pay attention to in their rating processes allows for clearer thinking regarding the construct(s) that the scores represent. Insights into the rating process can also extend understanding of the way various aspects of the rating context lead to score variations. Beyond issues of the

reliability of scores, results from this study will provide more insightful information about the role of each criterion in performance-based assessment, and the clarity and helpfulness of the rating scales and test validation. Second, a better understanding of how novice raters evolve into experienced raters should enhance the effectiveness of rater training programmes and the standardisation procedures since the detailed analysis of what scoring behaviours are associated with novice, developing and experienced raters is provided. These are the three overarching aims of the current research project (see section 1.3 for the stated aims).

Further understanding of the rating process is of more significance in the context where English is assessed by raters for whom English is not their first language. This is because raters' exposure to the first language (L1) of TTs may impact the rating quality and this effect may be stronger for speaking assessments (Knoch, Fairbairn, & Jin, 2021). Among the very few studies investigating this area is Zhang and Elder's (2011) study comparing L1 English and L1 Chinese in their ratings on the Chinese national College English Test - speaking component (CET-SET): the L1 Chinese raters were more concerned with language while the L1 English raters were more focused on content. However, it is inevitable that the findings will have been obscured by the interaction of other factors, such as the educational or training backgrounds of the raters involved. Elder and Davies (2006) proposed that there may be a rater distance effect, where languages more distant from the raters' own may be more difficult to rate, but according to Knoch et al. (2021) this is yet to be empirically shown in research on rating in language assessment. Therefore, there is a pressing need for an investigation into how English assessment is practiced in a context where English is a foreign language (EFL).

In light of these issues, this dissertation aims to extend our understanding of the rating process and, the factors that affect L1 Vietnamese raters' scoring

decisions, as well as their rating practice development, in the context of a speaking test of English proficiency. This research project does not aim to provide statistical measurements of rater behaviour; it aims to fill the gap by providing a rich description of the mental processes that the raters experienced, critically examining what may influence their rating decisions and how the raters have developed their rating practice over time.

1.2 Context of the study

A high level of education is not a sufficient condition to prepare oneself for challenges and competitions in the era of globalisation. According to the World Bank's survey on workforce skills in 2011-2012, competence in a foreign language was identified as one of the ten most important job-related skills in Vietnam (Bodewig, Badiani-Magnusson, Macdonald, Newhouse, & Rutkowski, 2014). English has a high social status in Vietnam due to the country's participation in regional organisations such as the Association of Southeast Asian Nations and Asia-Pacific Economic Cooperation, and the entrance of international organisations into the country such as the World Bank, and the World Trade Organisation (Phan, 2021). The Vietnamese government puts English at the heart of their language education policy as they consider English as the linguistic instrument for the nation to develop and modernise its economy, and participate in the global economy (Le, Nguyen, Nguyen, & Barnard, 2019). For Vietnamese people, English proficiency plays a significant role in achieving educational success, professional development and economic prosperity (Le, 2019). Therefore, in metropolitan areas of Vietnam such as Hanoi and Ho Chi Minh City, parents choose to provide their children with enhanced opportunities to learning English in English academies from a young age (H. T. M. Nguyen, 2011) and consider it an early investment (Le et al., 2019). For Vietnamese students from high school to tertiary education, the most popular aspirations in learning English are to fulfil the learning requirement at school, to

pass the national exams and to access better job opportunities (N. Nguyen, 2012) upon graduation. Other aspirations include pursuing postgraduate study or studying abroad.

Recognising the widespread importance of the ability to use English, Vietnam's National Foreign Language Project 2020 (NFL Project 2020) was launched with the aim of improving foreign language teaching and learning across the whole nation. The practice of English language assessments was the area of EFL policy that underwent the most prominent changes. The Common European Framework of Reference (CEFR), published by the Council of Europe (2001), has been a common concept to all Vietnamese English learners and teachers in recent years since the NFL2020 was initiated. NFL2020 was a huge and ambitious project with an aim to:

> *renovate thoroughly the tasks of teaching and learning foreign language, to implement a new program on teaching and learning foreign language at every school level and training degree, which aims to achieve by the year 2015 a vivid progress on professional skills, language competency for human resources […] By the year 2020 most Vietnamese youth whoever graduate from vocational schools, colleges and universities gain the capacity to use a foreign language independently. This will enable them to be more confident in communication, further their chance to study and work in an integrated, multicultural and multi-lingual environment.*
>
> *(Government of Vietnam, 2008, p. 1)*

The commitment of Vietnam to the globalisation process and its desire to boost the nation's competence in a foreign language are clearly seen in the level of English that it sets to different groups of learners at different year groups, as shown in Table 1.1 below.

Table 1.1: Graduation English language standard by level of education (Ngo, 2018)

Education level	*Target level (CEFR-VN)*	*Target level (CEFR)*
Primary	1	A1
Lower Secondary	2	A2
Upper Secondary	3	B1
Vocational Training	2	A2
Vocational Training (Advanced)	3	B1
University (Non-Major)	3	B1
University (Major)	5	C1
Community College (Major)	4	B2

Once the standards had been set, there was a pressing need for a national tool to measure the Vietnamese people's English capacity against these standards. However, before the NFL2020 project tests were often made by compiling different parts of available tests or resources from printed or online teaching and testing materials, or using the past papers of international tests such as IELTS, TOEFL, TOEIC, or Cambridge English Qualifications (T. N. Q. Nguyen, 2019). These tests, therefore, were mostly for institutional use due to the lack of validation research on the validity and reliability of the tests. This clearly showed the need for a national foreign language test, which led to the development of the Vietnamese Standardised Test of English Proficiency (VSTEP) as part of the NFL Project 2020. It was released nationally under Ministry of Education and Training Decision 729/QĐ-BGĐT on 11th March 2015 (2015).

The test is aligned to the CEFR and includes a multi-level test targeting levels B1-C1, as well as single-level proficiency certification tests, most notably at the A1, A2, and B1 levels. VSTEP test scores are important to TTs as it certifies people with a particular level of English proficiency, which can add or reduce credits to their academic and/or professional profile. English has been added as a requirement in job seeking, job promotion and even job security. For students,

VSTEP test results influence the placing of students in a particular programme or their graduation status. In light of this, the VSTEP is a high-stakes test since its scores are used to determine the futures of those who take the test.

While being aligned to international standards (CEFR), the VSTEP is a home-grown product since it was developed by Vietnamese language testers to assess L1 Vietnamese learners and users of English 18 years old and above. There are a number of aspects illustrating that the VSTEP is a localised test. First, the CEFR was contextualised to suit adults' contextual use of English in Vietnam. 75 skill-specific descriptors were added to the CEFR-VN version (see H. Nguyen, 2014; H. T. M. Nguyen et al., 2017 for detailed discussion). Second, the localisation of the VSTEP lies in its appropriateness to the education system of Vietnam (T. N. Q. Nguyen, 2019). Although the VSTEP is a non-curriculum related test, the set of themes, situations, and topics in the test reflect what Vietnamese learners learn in upper-secondary schools, colleges and universities. In addition, the test tasks were designed to echo the tasks and activities that Vietnamese learners perform at school (H. Nguyen, 2014). Third, the content of the test includes information about Vietnam or other Asian countries in order to maintain the TTs' interest and familiarity.

The VSTEP also falls into Glocal Type 2 tests as defined and categorised by Weir (2019) as it is localised in certain ways but global in others. Studies comparing VSTEP test scores with other international tests such as IELTS (H. Nguyen, 2014; Tran et al., 2015) and APTIS (Dunlea et al., 2017; Dunlea et al., 2018) provide empirical evidence of the equivalence of VSTEP test results with those from renowned international standardised tests. Since the VSTEP is the first standardised test "made in Vietnam" (T. N. Q. Nguyen, 2019, p. 78), further research is clearly needed to enhance the qualities of the test. T. N. Q. Nguyen (2019), a key researcher in developing the test, suggested that one of the top

priorities should be “to ensure its scoring validity through improving the quality of VSTEP speaking and writing raters” (p. 95).

The VSTEP consists of sections assessing reading, writing, speaking, and listening, with all four sections taken by all TTs. The reading and listening papers come in the form of multiple-choice questions, i.e., the TTs provide answers to the questions by selecting appropriate options A, B, C or D. The writing and speaking papers require TTs to provide written and spoken responses to a task, as defined by Davies et al. (1999, p. 196):

> *A type of test item involving complex performance in a test of productive skills. Examples of writing test tasks are: writing an essay, drafting a business letter, and completing a summary of a text. Examples of speaking test tasks are: adopting a specific role in a role play, describing a photograph, and presenting an argument to a small group of peers.*

However, such performances are more complex to score than multiple-choice items. Typically, such writing or speaking performances are evaluated by one or more raters who refer to rating criteria when making their decisions. Raters are therefore central to arriving at a score that summarises the performance and this score later forms the basis for making decisions and drawing inferences about the TTs. Thus, the reliability and generalisability of the scores from VSTEP writing and speaking tests have raised considerable concerns from TTs and other stakeholders. This study focuses on the speaking component of VSTEP tests.

VSTEP speaking tests consist of three parts (see Appendix 1 for a sample of a test). In Part 1 - Social interaction, the TTs are required to answer 3-6 questions on two different topics. This part lasts approximately 3-4 minutes. In Part 2 - Situation, the TTs are given a situation with three options to select. The TTs are required to select the best option and explain the reason(s) why they have chosen that option in preference to the other two. The TTs have 3-4 minutes to explain their choice. The third part of the test, which lasts 3-4 minutes, provides the TT with a topic and a mind map of suggested ideas of how to develop the

given topic. The TTs can use the suggested ideas or his/her own ideas. If time allows, the TTs discuss several follow-up questions upon completion of Part 3 (Topic development). In live test administrations, TTs' performances are examined on five criteria in the ten-point rating scale by one rater. The criteria comprise: Grammar, Vocabulary, Pronunciation, Fluency and Discourse management. The scale aims to examine test takers' proficiency levels at B1, B2 and C1 according to CEFR levels. An example of the rating scale is provided in Appendix 2; however, due to authorised restricted access, only descriptors of the first three band scores of the five criteria are available in the appendix.

Several studies into the rating quality of VSTEP speaking tests (T. N. Q. Nguyen, Nguyen, Nguyen, & Carr, 2017; T. N. Q. Nguyen et al., 2020) reveal that although the majority of VSTEP raters were consistent with their ratings (T. N. Q. Nguyen et al., 2017), the raters exercised considerably different levels of severity. Rater training programmes were shown to reduce raters' effects and bring raters into better alignment; however, a small number of raters who were outliers persisted (T. N. Q. Nguyen et al., 2020). One shortcoming of these studies is that they relied heavily on quantitative measurements. More qualitative studies can provide better insights into why VSTEP raters behave the way they do in their ratings, thus providing a basis to enhance the effectiveness of rater training and therefore rating quality. Thus, the current study focuses on these issues as this contributes to the originality and significance of this research project in Vietnam's context in particular and to rating literature in general.

1.3 The aim of the study

The investigation of this study involves three overarching aims:

- To further an understanding of the rating processes experienced by Vietnamese teacher-raters in the assessment of speaking performance in English as a foreign language
- To further an understanding of the factors which affect the raters' scoring decisions
- To further an understanding of the ways raters develop their rating practice over time

To this end, I critically reviewed relevant literature in order to establish the state of the existing knowledge of oral language rating in a high-stakes testing context. I then conducted an empirical investigation with the goal of generating and interpreting data and possibly advancing the knowledge base relating to oral rating in a high-stakes language testing context.

1.4 Research questions

In order to achieve the stated aims, this research project addresses the following questions:

1. What are the mental processes of rating speaking performances?
    - What differences between experienced and novice raters, if any, are seen in attention paid to language features described in the rating scale?
    - What differences between experienced and novice raters, if any, are seen in decision-making strategies?
    - What differences between experienced and novice raters, if any, are seen in attention paid to test-taker proficiency levels?
2. What are the factors that cause disagreements among all the raters in making score decisions?

- In what ways and to what extent do these factors affect raters' decision in their ratings?

3. In what ways do raters develop their rating practice?

1.5 Significance of the study

There are important areas where this study makes an original contribution to the context and to the literature. First, with growing concerns of the validation of the VSTEP tests, there is a lack of understanding of the rating process experienced by VSTEP speaking raters, what factors can impact their rating decisions and how VSTEP raters develop their rating practice. The results of this study can extend our understanding of these issues, thus providing valuable information to test developers and policymakers to improve the quality of the test within Vietnam's education context. The findings from this study could have important implications for developing and revising VSTEP rater training and could accordingly enhance the quality of ratings. In addition, this study could form a valuable source for future research projects about VSTEP speaking and VSTEP rater training for the study context.

In terms of literature, this study will contribute to the debates about the complexity of the rating phenomenon from an under-researched context, i.e., that of a Southeast Asian country. The practice of assessing English should be investigated in different educational cultures and backgrounds (Sheehan & Munro, 2017) in order to further our understanding of the complexity of rating in a local and global context. It is important to address the local and global dimension of the English language in high-stakes language examinations (Sewell, 2013) and in local testing contexts (Dimova, 2017; Harding, 2017), although Sewell (ibid) accepted that the task of negotiating the two dimensions is a challenge to language testers (more discussion in section 2.8.2). Thus, this study is valuable since it adds new insights to the literature about the practice of

assessing spoken English in a local context by its local people for whom English is not their first language. Furthermore, rater scoring has an impact on the reliability and validity of performance-based tests. Thus, issues related to rating have received a continued call for further investigation to extend understanding of the phenomenon (Crusan, 2015). Myford (2012) urged researchers and test designers to "do all that we can to help ensure that the ratings that raters assign are accurate, reliable, and fair" (p. 49). This project provides an important opportunity to advance understanding of the nature of the interaction between different factors that could mediate the oral rating process and rating practice development.

1.6 Reasons for selecting the topic

The reasons why I selected the thesis topic were my interest in language education research and my experience of working within the field of language testing and assessment. I worked for several years in the field of ELT as a university lecturer and a teacher trainer. I have witnessed the changes of language education policies in Vietnam in general and in language assessment in particular. The adoption and adaptation of the CEFR was a major language education reform, which has resulted in considerable changes in how English is taught, learnt and assessed in the country. In 2012, several key members of the Institute where I worked (including myself) were given a valuable opportunity to study a professional development course in Language Assessment in Sydney, Australia. The course helped us further our understanding of the CEFR and its potential application to syllabus design and assessment. Upon completion of the course, we returned to our Institute and together applied what we had learnt to make necessary changes in our language education courses and assessments. All courses were re-designed to align to the CEFR. The Institute was one of the first in a network of language education institutions in the country to try to

incorporate the CEFR into its courses and align assessment tasks to the CEFR. I was a leader in charge of developing two courses which aligned to B2 and C1 levels. I was also responsible for developing and revising end-of-term tests of English for undergraduates in their second year. Thus, my interest in the field of language testing and assessment, and participation in the implementation project, can be traced back to my work at the Institute.

My interest in the field increased further when I was invited to join another team from the Institute to work on VSTEP tests. As discussed in section 1.2, one of the innovative reforms in the nation's language assessment was the development of the VSTEP test – the first ever standardised test of English in Vietnam. There was an increasing need to enhance the test quality to gain trust from different stakeholders and the public. I was involved in writing test items, running statistical analysis of the test data to revise the test items and enhance rating quality. I was also a trainer in the VSTEP rater training and monitored the rating procedures. In section 3.10 is a discussion of how my roles in the Institute and in this team may have influenced this current study. Rating quality, therefore, played an important role in my work experience. I was intrigued by my discussions with active trainees and raters about the rating procedures. Although statistical analysis could inform me about their rating behaviour (e.g: (in)consistency, reliability, severity), the underlying reasons for the behaviour were a mystery to be understood. As this work will be submitted for the Degree of Doctor of Philosophy, a degree for professionals, I was keen to focus on a topic which is of immediate concern in my professional life.

Another reason for choosing this topic was that I seem to have found a gap in the research literature. As will be explored in more detail in chapter 2, I discovered that there was little research into the mental processes of rating speaking performances in a high-stakes test and into the interaction of the

factors which can affect the processes and the rating practice development. I became further convinced that the topic was worthy of investigation when I was accepted to deliver papers at several conferences (e.g., LTF, 2018; AALA, 2019; LTRC, 2021) and received the best poster prize at LTF 2018, the best student paper at AALA 2019, the British Council Assessment Research Award in 2019 and the Santander research support grant in 2021. The positive feedback confirmed to me that the topic would be suitable for this dissertation. A list of papers presented and awards received is provided in Appendix 3.

1.7 Definitions

The use of terminologies in the published literature on rater behaviour is sometimes ambiguous. For example, norming and training are both used to refer to activities to familiarise raters with the rating scale. Therefore, in order to enhance clarity, definitions of a few key terms, which were adapted from Davis (2012, p. 8-9), are given below.

1.7.1 Rater behaviour

Rater behaviour refers to the behaviours which can be observed while raters are rating, such as interactions with the rating scale and/or the performance and the time taken to reach a scoring decision.

1.7.2 Rater cognition

Rater cognition refers to the processes occurring in raters' minds during scoring. This includes both the attention raters pay to the features of test takers' performance while scoring, and the mental actions taken to arrive at a score.

1.7.3 Rater training

Rater training (sometimes 'rater training programmes') refers to activities undertaken by raters that are intended to enhance rating quality. Training also refers to activities occurring outside of operational scoring.

1.7.4 Rating experience

This thesis adopts the definition by Lim (2011) that experience refers to raters' previous participation in scoring activities within a specific testing context, the VSTEP test. It does not include aspects of the rater's background or the rater's previous experiences outside of the testing context.

1.7.5 Rater severity/leniency

Rater severity refers to the tendency of awarding lower scores while another awards higher scores for the same performances. Rater leniency refers to the tendency of awarding higher scores while another awards lower scores for the same performances.

1.8 Structure of the thesis

The thesis comprises seven chapters.

Chapter 1 – Introduction outlines the background context of the study, the area under research, the overall objectives and provides an overview of the study and organisation of the dissertation.

Chapter 2 – Literature review critically analyses empirical studies relevant to this research.

Chapter 3 – Methodology presents the methodological approach taken in this study and provides a rationale for the methods and procedures employed throughout the data collection and analysis processes.

Chapter 4 – presents analysis and findings in relation to the scoring process as experienced by novice and experienced raters.

Chapter 5 – presents analysis and findings in relation to the factors causing disagreement among all the raters.

Chapter 6 – presents analysis and findings in relation to the ways raters develop their rating practice.

Chapter 7 – Discussions, Conclusions and Implications summarises the study's findings and posits some contributions it makes to the field. It provides comments upon the strengths and limitations of the study, as well as its potential contributions to English speaking testing.

Chapter 2: Literature Review

2.1 Introduction

In language performance tests, raters are important as their scoring decisions determine which aspects of human performance the scores represent; however, raters are considered as one of the potential sources contributing to unwanted variability in scores (Davis, 2012). Therefore, the last three decades have seen a growing interest in trying to understand the variability in raters' scoring patterns and the factors which contribute to this variability. This chapter reviews issues surrounding the variability of rater judgements in language tests and how rater-related factors may influence rating quality. Section 2.2 traces the history of language performance-based assessment, highlighting the significance of enhanced rating quality. Section 2.3 of the chapter discusses two models proposing factors that may impact the rating quality: one suggested by McNamara (1996) and the other suggested by Knoch et al. (2021). The focus of this research project is the mental processes of rating and the factors influencing this process; therefore, sections 2.4, 2.5, 2.6 and 2.7 explore in detail the issues of how rater background, rater experience, rater cognitive process and a community of practice may influence scores and decision making. Section 2.8 follows with an examination of the influence of different perceptions of the rating criteria, including fluency, pronunciation, lexis, syntax and discourse competence aspects. This chapter ends with implications of the literature review for the current research project in section 2.9 and argues that I have discovered a gap in the literature. It is important to note that the chapter includes discussion of a small number of studies which were conducted in the 1990s (e.g.: Cumming, 1990; Lumley and McNamara, 1996; McNamara, 1996; Vaughan, 1992; Weigle, 1992; 1998) as the findings of these studies are important and influential on later work in that they established key issues related to the topic of this dissertation. Moreover, the age of these studies also demonstrates that

investigation of this topic has been a concern for a long time, yet a sound answer has not yet been revealed (more detailed discussions are presented in later sections of the chapter). Additionally, even though distinctions do exist between the way writing tests and speaking tests are marked, the effects of training sessions, rater experience and raters' interaction with scales on the raters of writing tests can provide essential insights for research concerning raters of oral proficiency. Therefore, such studies are occasionally discussed in this chapter.

2.2 Historical perspectives

This section traces the history of performance-based rating and presents the two following important points:

- The difference between traditional fixed response assessment and performance-based assessment
- The significance of adequate understanding of fair measures of the test takers' ability in performance assessment settings

The twentieth century witnessed one of the most prominent changes in education assessment as traditional test formats (i.e.: pencil-and-paper tests involving multiple choice questions) were increasingly replaced with performance-based assessment (McNamara, 1996). The driving force behind these changes had been government policy, which required learners to demonstrate practical command of skills in all areas of education (E. Baker, 1995). Language assessment was no exception to this move as language plays a crucial role in the workplace.

McNamara (1996) provided a representation of the features of a typical second language performance test in comparison with a traditional pencil-and-paper language test (adapted from Kenyon, 1992), (Figure 2.1):

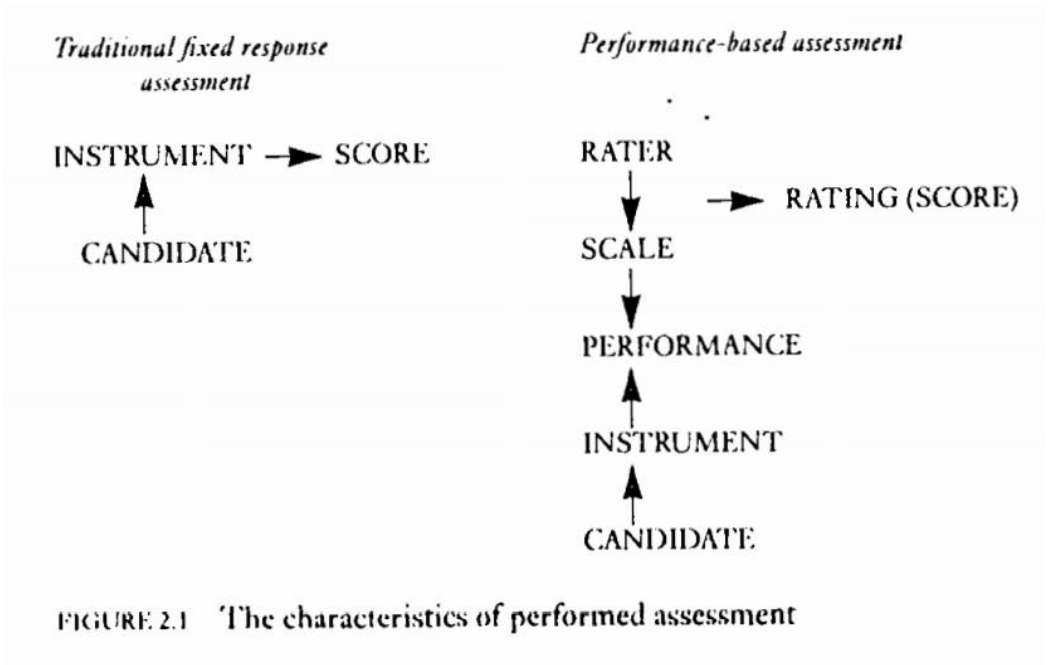


FIGURE 2.1 The characteristics of performed assessment

Figure 2. 1: The characteristics of performed assessment (McNamara, 1996)

The traditional fixed response assessment provides the test taker with an assessment instrument which contains a number of options, for example A, B, C or D, only one of which is correct. The scoring task is simply to count the correct responses, which are transparently and objectively indicated by a checked box or a circled number. Thus, the score is the direct result of the instrument since possible responses from the candidate are anticipated in the form of the instrument itself. In performance-based assessment, the candidate interacts with an instrument to produce a performance (such assessment is often termed constructed response assessment) instead of selecting available choices. The human rater then scores or judges this performance by using a scale or other kind of scoring schedule. The interaction between the rater and the scale is a new type of interaction which arbitrates the rating process. Typically, the two features that distinguish a performance-based test from a traditional fixed response test are "a *performance* by the test taker which is observed and judged using an agreed *judging process*" (McNamara, 1996, p. 10).

The assessment of the performance (written or spoken samples of language) created by the test taker is almost impossible to measure objectively with human judgement. The psychometric view of assessment was a challenge for language testers who wanted to test performances in 1930s and 1940s (Spolsky, 2017). On the one hand, language testers were challenged to establish the reliability of their judgements and on the other hand they needed to determine the factors leading to individual judgement. The very first attempt to assess speaking performance was seen in the Foreign Service Institute Oral Proficiency Interview (O'Sullivan, 2012) when the Assistant Secretary of State insisted that language proficiency (speaking ability included) of American diplomats be tested (Spolsky, 2017). This test was also the first to develop a system of testing using two or three judges and a scale (Spolsky, 2017). This became the model for such testing in other government agencies and later in a wide variety of educational settings, including foreign language testing (L. Bachman & Palmer, 1981) .

The scale is an increasingly important area in the field of testing, as stated by Spolsky (2017). One of the first scales in the testing history which was developed by Thorndike consisted of a set of exemplar scripts of handwriting being judged. The set of exemplar scripts was selected from a large number of handwritings and ranked by 200 teachers (Thorndike, 1910). Another scale used by the Foreign Service Institute was a set of descriptions that particularly describe the characteristics of language performance at a certain level or stage of learning (Jones, 1979). Such scales reminded the trained judges of the standards prescribed in their training sessions and of their previous experience.

There continued to be pressure since performance assessment involves judgements of quality against such rating scales; new features of the assessment setting were introduced such as:

(1) The raters themselves, who will vary in the standards they use and the consistency of their application of those standards;

(2) The rating procedures they are required to implement.

(McNamara, 1996, p. 3)

These new features have major implications for research in the rating process as the complex interaction between rater characteristics and the qualities of the rating scales appear to strongly influence rating quality, regardless of the quality of performance. In order to enhance reliability of scores, McNamara (1996) argued that it is necessary to investigate and control for the effect of rater variation and scale characteristics.

One of the earliest studies investigating the reliability of ratings of second language assessment was Coffman's (1971), which summarised observations of how raters differed from each other, including:

(1) Different raters tend to assign different grades to the same paper

(2) A single rater tends to assign different grades to the same paper on different occasions

(3) The differences tend to increase as the essay question permits greater freedom of response

(Coffman, 1971, p. 26)

Later in 1990, L. Bachman, one of the most influential figures in the field, introduced his model of language ability and demonstrated how this could be used to support a language test. He mphasized the significance of identifying sources of error and estimating the magnitude of their effects on test scores. He also proposed ways to increase the reliability of measures. In terms of scoring, he suggested several statistical techniques (e.g.: coefficient alpha) to measure rater consistency, including reliability within a single rater (intra- rater reliability) and among different raters (inter-rater reliability). In order to extend understanding of rater behaviour, the use of multi-faceted Rasch measurement

developed by Linacre (1988) started to gain popularity and has been used widely in language assessment related research (McNamara & Knoch, 2012).

Spolsky (2008) draws attention to what he considers mistaken steps in our past, that is:

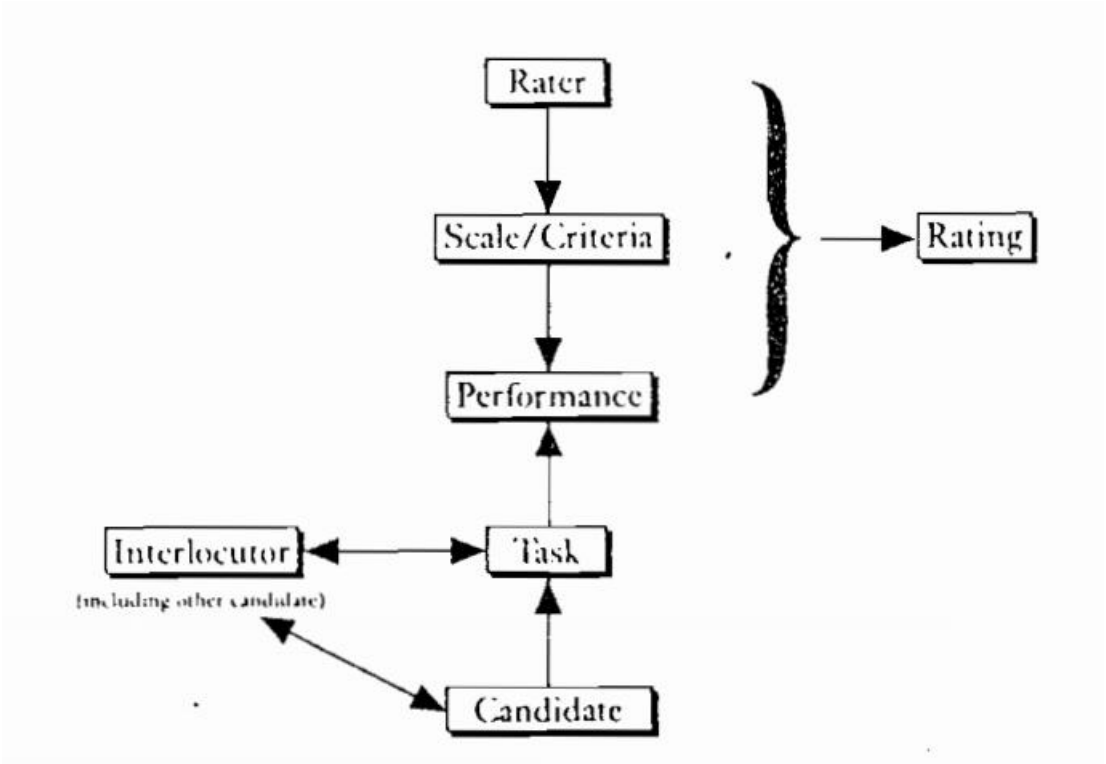
> *"When we realized over 100 years ago the inevitability of error in the measurement of human capacity (Edgeworth, 1890), we set out to try to reduce the size of the error, rather than trying to understand the risk of making decisions about the fate of human beings using erroneous data"*
>
> *(Spolsky, 2008, p. 302).*

In other words, identification of rater errors from a psychometric perspective provides little insight into the reasons they occur. As stated in chapter 1, the rating process is still not greatly understood, and there have been numerous calls for the expansion of research initiatives in this area. Thus, it is important to extend the understanding of how scoring decisions are made by human raters so that the insights can inform the meaningfulness of the scores awarded. Two models suggesting the factors influencing rating quality in rater-mediated assessment are discussed in detail in the next section.

2.3 Models of factors influencing rating quality in rater-mediated assessment

The earliest model suggesting the potential influences on rating quality was proposed by McNamara (1996) (Figure 2.2).

Figure 2. 2: Performance-based assessment (McNamara, 1996)



Instead of placing emphasis on the candidate ability and assessment tasks as the only factors which affect the scores, the model shows that language performance assessment is mostly rater-mediated. This means a TT may have received a different score for the same performance if s/he had had a different rater. Moreover, the age, gender, educational level and personal qualities of the interlocutor may influence the candidate's performance. This model raised awareness of the need for better understanding of these variables among raters. This consideration does not aim at blaming individual raters because errors may result from the rating procedures, or the rating instruments. It does, on the other hand, aim at producing the fairest scores for the test takers. The model, however, is basic and does not present a detailed explanation of the rating process in which the ratings operate and the complex factors that come into play.

25 years later, being aware that raters bring with them a variety of experiences, values and backgrounds into their ratings, Knoch et al. (2021) proposed a more detailed model illustrating a number of aspects influencing rating quality (Figure 2.3) by expanding McNamara's (1996) model.

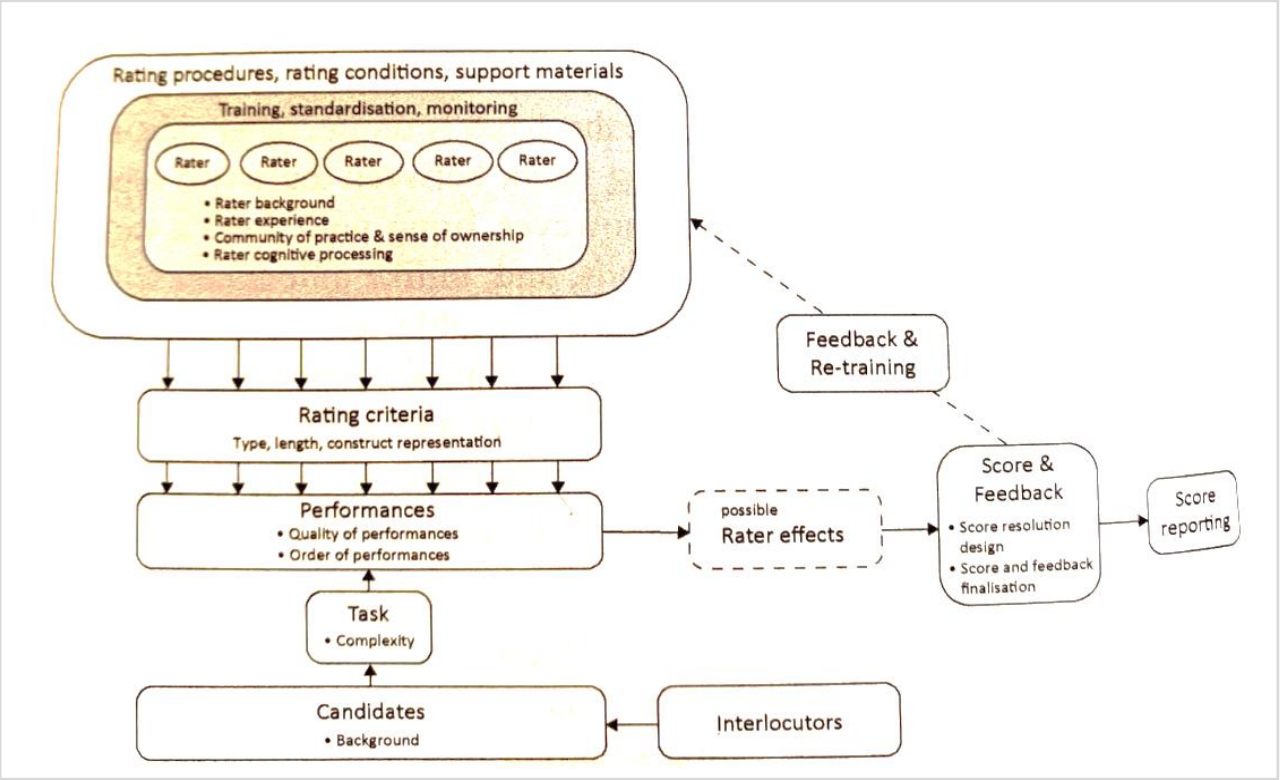


Figure 2. 3: Factors influencing rating quality in rater-mediated assessment (Knoch et al., 2021)

As can be seen in Figure 2.3, the raters as a group are influenced by the rating procedures used, the rating conditions in which they work, the rater training and the rating support materials available, as well as any standardisation and monitoring they experience as part of their rating practice. The model suggests that rater background, rater experience and rater cognitive processing are among the factors which the raters bring with them into their ratings. Moreover, the authors depicted the raters as a community of practice to show that this assessing is often done within an environment which provides a sense of a group, a common goal and an opportunity for conversations and discussions.

This sense of a community of practice may also affect rating behaviour as may rater cognitive processing. In addition, the raters, together with the four immediate factors (depicted in the inner grey box), interact with the rating criteria to evaluate spoken and written performances. This interaction may influence the raters' decision-making and therefore their scores. Raters may also be affected by certain candidate characteristics or the perceived task complexity, as well as the performance of any interlocutors involved in the performance. In this research project, rater-related factors, which are illustrated in the inner grey box, together with their perceptions and application of rating criteria are examined in detail to fulfil the aims of the project.

Knoch et al.'s (2021) model was chosen due to the complex nature of rating speaking performances and the different related factors affecting construction and the process of rating in the context of this study. It promotes interrogation of the participants' backgrounds, teaching and rating experience, perceptions and consideration of how these factors affect their ratings. Over the last three decades, extensive research studies on rating quality have been conducted to explore the complex phenomenon of rating quality, highlighting different influential factors. To name but a few, A. Brown (2000, 2006) examined what IELTS raters attended to during their ratings, while Harding (2016) studied the usability of a pronunciation rating scale. However, these studies are mainly exploratory, investigating only one or two factors at a time. McNamara's (1996) model was not chosen for the current research due to its limitation to fulfil the aims of the research project. This basic model of factors influencing rating quality does not present a sufficiently detailed explanation of the process in which the ratings operate and the complex factors that come into play.

It is important to note that Knoch et al.'s (2021) model was only published recently while this thesis has been in development since 2017. The previous

version of this literature review chapter was similarly structured as it also discussed the factors that potentially impact the rating process and the raters' scoring decisions, including rater cognition, teaching experience, raters' linguistic background, rater training, and raters' perceptions of rating criteria. Knoch et al.'s (2021) model helps to confirm the significance of these variables in understanding the research phenomenon and highlights the need to examine how these key factors might differ and under what circumstances. Moreover, the model helps to identify an additional factor that is relevant and important in explaining the meaning associated with the researched phenomenon, that is the community of practice (see section 2.7 for more details).

The following sections of this chapter are, thus, organised according to the factors suggested in the model of Knock et al. (2021). Section 2.4 discusses how rater background can contribute to score variations. Sections 2.5 and 2.6 examine in detail the influence of rating experience and rater cognition on decisions of scores, respectively. The way raters interact and develop their community of practice and how being in this community may impact their rating is discussed in section 2.7. Section 2.8 explores how different perceptions and measurements of rating criteria can lead to differences in rating decisions. As the focus of this research project is on the rating processes and rating practice development, other factors in the model are not discussed.

2.4 Rater background

2.4.1 Teaching experience

EFL/ESL teachers are often recruited to be exam markers in both international proficiency tests such IELTS, Cambridge tests and national tests such as Jitsuyo Eigo Gino Kentei/Test in Practical English Proficiency (EIKEN), and the VSTEP. There are a number of benefits of teacher involvement in high stakes tests. By

becoming an examiner, teachers develop their own assessment literacy (Goldberg, 2012) by having first hand insights into explicit assessment standards (Klenowski & Wyatt-Smith, 2012), thus regulating their instruction in classroom to be clearer (Buck, Ritter, Jensen, & Rose, 2010) and allowing them to make more precise judgment about students' performance (Klenowski & Wyatt-Smith, 2012).

Teacher involvement, without the necessary standardisation, in a national examination, also poses several threats because potential variability in rater judgments can contribute to measurement error (McNamara, 1996), particularly when teachers bring their own beliefs in interpreting assessment criteria (Cheng, 2008). Davison (2004) argued that high stakes tests should take into account teacher raters' interpretation of the assessment criteria as different interpretations may exist due to the difference in teacher raters': "personal background, previous experience, unconscious expectations, internalised and personalised preferences" (p. 308). More importantly, high stakes tests may provoke teachers' negative feelings when teachers' beliefs are not parallel with the standards set by the test (Costigan, 2002). Therefore, further studies providing greater insight into the impact of teacher beliefs on their rating decisions are of significance (Davison, 2004).

Among the very few studies investigating the direct impact of teacher beliefs on their scoring behaviours in high-stakes national tests, Hsieh's (2011) and Goh and Ang-Aw (2018)'s studies are two of the latest studies. Hsieh (2011) examined the differences between ratings of linguistically naïve undergraduates and ESL teachers using a holistic proficiency scale and a scale focusing on accentedness and comprehensibility. The teachers were found to be more lenient on these two aspects but there were no statistical differences between the ratings of the two groups on overall proficiency. An analysis of the written

comments provided by the teachers, however, showed that the lay judges focused more on global qualities whereas the teachers focused on more specific qualities in the performances. These findings seem to indicate that the teachers might have brought their ESL teaching experience into their ratings. What the study lacks is more qualitative data to gain a deeper understanding into the decision-making processes of raters and how they relate to the construct of the assessment.

In Goh and Ang-Aw's (2018) study, the issues of how teacher-raters applied their beliefs and the descriptors in their ratings were examined through the analysis of the verbal protocols of seven raters who were experienced teachers. Although all the raters received the same training on the rating process, the findings revealed differences in raters' perceptions of personal responses, speech-related considerations, test taking strategies, and their application of the rating scale. For example, a global impression of performance played a role in the raters' evaluation; however, it is not an assessment criterion mentioned in a rating scale or the questionnaire. The findings would have been more helpful if the authors had provided more details on the differences of teachers' teaching experience and speculated how such differences could lead to differences in their rating behaviours.

While Hsieh's study was conducted in the context of international teaching assistants in the US, Goh and Ang-Aw studied the ratings of the 'O' level oral national examination in Singapore. In a slightly different context – teacher-based assessment in secondary schools, Davison (2004) compared the different assessment beliefs, attitudes and practices of two groups of ESL teachers in evaluating ESL essays of senior secondary Cantonese-speaking students in Australia and Hong Kong. There were different assessment beliefs among these teachers, and this consequently led to different rating behaviours in Australia

compared with Hong Kong. It was found that there were two conflicting assessment approaches Australian teacher raters used. One approach was "a legalistic process of ticking boxes, while at the same time recognizing the ethical dilemmas created by their own 'humanity'" (p. 319). By way of illustration, while the Melbourne teachers valued the criteria as a way of 'keeping them honest', they also emphasised their relationship with and their knowledge of the test takers when making their score decisions. In contrast, it was revealed that underlying assessment criteria seemed to be the cause of variability in the Hong Kong teachers' judgment. For example, some schools traditionally focus more on accuracy and mechanics while others – because of their underlying philosophy and student profile – give more emphasis to creativity and content. Because of this difference, some Hong Kong teachers were focusing more on grammar and sentence level structure and others more on ideas and overall organisation. Davison (2004) suggested that the variability in Hong Kong teachers' judgement might have been due to the lack of agreed published criteria and the lack of widespread teacher training. The study concludes that traditional notions of validity may need to be reconceptualised in high stakes teacher-based assessment, with professional judgment, interaction and trust given much higher priority in the assessment process.

The studies reviewed above have brought into focus teachers' beliefs and practices in the literature of assessing speaking performance in high stakes tests. Vietnam's context is, to some extent, similar to Hong Kong's in the way that different priorities towards speaking proficiency indicators may exist because of different demands on different groups of learners and different understandings of the curriculum and assessment goals. All VSTEP raters are EFL teachers; thus, I would argue that it is important to unpack teacher-raters' belief systems and practices in relation to their rating processes because it will provide insightful information for test administrators, policy makers and other stakeholders in

Vietnam about the meaning of scores awarded by VSTEP teacher raters. As Davison (2004) argued in her study, "even when it is possible to establish common understandings of the task, high-quality publicly agreed and explicit assessment criteria, and strong moderation discussions, teacher interpretation [of the rating criteria] will always be needed. This should be seen as a strength and not a weakness of teacher-based assessment" (p. 328).

2.4.2 Linguistic backgrounds

Linguistic experience in terms of accent familiarity tends to enhance listeners' understanding of particular accented speech (Harding, 2012; Major, Fitzmaurice, Bunta, & Balasubramanian, 2002). However, researchers in language assessment have considered this as a potential source of rater bias in their evaluation of speaking ability. Thus, raters' familiarity with TTs' first language has been an active area of investigation for researchers in oral assessment. In this study, I focus on exploring the situation where raters judge the speaking proficiency of TTs who share the same L1 as them, because those studies will likely best inform my research project in which Vietnamese raters are employed to examine Vietnamese people's English proficiency level in the VSTEP test - a locally developed test. I look particularly at whether Vietnamese raters' scoring decisions are influenced by their preference for and familiarity with the TTs' accents. Thus, in this section, studies investigating the effects of local raters' familiarity with TTs' L1 are thoroughly discussed.

To understand how ratings are influenced by raters' familiarity with test takers' accents, Caban (2003) examined differences in the rating behaviours of 4 groups of raters (ESL trained L1 English, EFL trained L1 Japanese, L1 Japanese and L1 English) when evaluating four responses of Japanese accented speech. Statistical analysis of the data showed that EFL-trained Japanese L1 raters were more severe in evaluating pronunciation and grammar. Two categories focused on EFL

settings in Japan while both L1 English groups were consistently lenient on pronunciation. The author provided a quite convincing explanation that in Japan, the English curriculum prioritised the ability to control pronunciation and grammar; thus, these raters may have higher expectation of the TTs' pronunciation and grammar, leading to harsher evaluation of these aspects. This, to some extent, led me to have some initial thoughts that the perceptions raters hold can affect their scoring decisions, which seems to be supported by Goh and Ang-Aw's (2018) study reviewed in the *Teaching experience* section. The findings of Caban's (2003) study would have been more convincing if the raters had been required to score more varyingproficiency-level performances, since a greater number of ability levels could have given more insight into rater behaviour. Moreover, since the raters were not trained before performing their ratings, the raters may have behaved more differently than if they had been trained. Another issue that may render the findings questionable was the use of a 15-point rating scale on 7 categories. According to Xi and Mollaun (2009), this may be because raters would find it challenging to differentiate among the scales and evaluate all 7 of the criteria at the same time.

To further explore this issue, Xi and Mollaun (2009) looked at whether a special training on accent familiarity can lead to differences in the local (in this case Indian) EFL teacher raters' scoring decisions. It was found that raters with special training did not outperform those with regular training at the task level; however, statistical analysis seemed to show that raters in the special training group marked Indian examinees' performance more consistently than raters in the other group. The reason for this consistency was thought to be the raters' familiarity with various (Indian) accents, which helped the raters better understand Indian accented responses. However, the authors did not think raters' linguistic backgrounds contributed to raters' leniency. Since the

interpretations were speculative, the authors called for qualitative research to cast light on their interpretations of the findings.

Another study (O. Kang, Rubin, & Kermad, 2019) sought to estimate which variables related to rater background have an impact on novice raters in their assessments of L2 spoken English. Among the variables investigated, the status of L1 English speaker was the most salient across the analyses. Raters for whom English was not their first language were more severe than L1 English raters. One possibility to explain this rating behaviour was provided by the authors. It might have been due to the different attitudes non L1 English raters had towards English varieties and the inner-circle English pronunciation target (Kachru, 1992) which they themselves have struggled to acquire. That speculation, however, remains in need of empirical testing, as suggested by O. Kang et al. (2019). Moreover, as the study recruited eighty-two untrained raters who were undergraduates, the findings seemed to be self-explanatory.

The three studies above attempted to examine whether raters' L1 language and the L1 of TTs were related through scoring but all of the studies used holistic rating scales. According to Knoch (2009) and Li and He (2015), raters behave differently as a result of using holistic or analytical rating scales. It can be assumed in these studies that raters who shared test takers' L1 may be unable to resist their preference towards accent familiarity during the rating process and that such preferences will contribute to a more lenient rating of overall performance, not for individual criteria. Hence, more research is needed to see whether such differences persist when an analytical rating scale is used. Furthermore, all the previously mentioned studies suffer from shortcomings as they relied too heavily on quantitative analysis.

While Caban (2003) and Xi and Mollaun (2009) used quantitative methods to analyse the collected data, Y.-H. Kim's (2009) study contributed to a better

understanding of the issue by using both quantitative and qualitative methods to investigate evaluation behaviours of 12 English-L1 teachers of English and 12 Korean-L1 teachers of English in assessing 10 Korean-accented speeches. The overall results of three different statistical approaches indicated that little difference was found in internal consistency and severity in both groups of teacher raters. However, the qualitative analysis of raters' written comments showed that the two groups of teachers sometimes did not pay equal attention to the same rating categories. For example, accuracy of information was the focus of English-L1 teachers' concerns whereas the way information was delivered by the test takers received more attention from Korean-L1 teachers. One of the limitations with this explanation is that the rating scale, the criteria and the training were not described in detail, so the teacher raters may not share the same understanding of the requirement.

Another mixed method study conducted with 19 English-L1 teachers and 20 Chinese-L1 teachers of English (Zhang & Elder, 2011) revealed similar results to Y.-H. Kim's (2009) study , i.e., there was no statistical differences in the scores assigned by the two groups. However, qualitative data suggested that Chinese-L1 raters appear to pay more attention to language form and less to communication than English-L1 raters. This seems to contrast with recent descriptions of World Englishes or English as a lingua franca (ELF) communication, where it has been argued that getting the message across is what matters for ELF learners and/or ELF users rather than approximation to L1 English based models (Jenkins, 2006). Furthermore, Zhang and Elder (2011) did not suggest that Chinese raters were operating according to a different code, instead the researchers speculated that their participant raters may have been more oriented to standard English than the English-L1 group. This finding is important as it draws out the need for further studies in a context where English is not the first language used. In particular, more qualitative research is needed

to unpack how and to what extent standard English or English varieties influence scoring decisions of English-L2 raters in contexts such as China and Vietnam, which is a less researched area.

All of the studies reviewed above varied one from another in terms of the raters being trained or untrained, with or without EFL/ESL teaching experience, holistic or analytical rating scales and the assessment contexts, which leads to difficulties in comparing the findings of these studies. However, in this section I present evidence that there was no statistical difference in ratings between L1 English and L2 English raters. I suggest that additional qualitative studies will be needed to develop a fuller picture of the influence of raters' linguistic background on their ratings. Particularly, I propose that it is important to investigate how EFL trained raters perceive and perform ratings using analytical ratings scales in their local context, and whether they are unconsciously applying a L1 English standard at the top of the scale (Hill, 1996, p. 45) because of their linguistic preference/background.

2.4.3 Rater training

Research into the effectiveness and limitations of training sessions on rater behaviours suggests that rater training sessions have an influence on raters' decision-making behaviours in different ways (Barrett, 2001; Eckes, 2008, Knoch, 2010; Vaughan, 1992; Lumley & McNamara, 1995; Weigle, 1994, 1998). These influences are outlined below.

There appears to be a trend that training sessions do not succeed in achieving the aim of minimising the differences among raters' behaviours. For example, an analysis of the verbal reports of nine experienced raters in Vaughan's (1992) study revealed that raters slid away from the guidelines for essays which did not fit the descriptors of the holistic scale even though they had reached a

consensus on many essays in the training given before. In a similar vein, evidence from Barrett's study (2001) indicated that considerable differences in rater severity or leniency and a lack of agreement between raters existed among sessional staff, and non-regular raters. One major strength of this study is the use of the Partical Credit Model – an extension of the Rasch model – to analyse the performance of raters before and after the training session, since a clearer picture of how the raters interacted with each other and with the items was depicted with statistics. However, no attempt was made to uncover the sources of such rating errors.

On the other hand, it has been suggested that training sessions helped improve rating reliability among four novice raters in different positive ways including the shared understanding of the rating scale, and the self-awareness of their ratings compared with other raters (Weigle, 1994). In her later study, Weigle (1998) explored the effectiveness of training sessions on a new aspect of rater behaviour: severity and consistency in giving scores. The results suggested that training seemed to reduce the extremism of the new raters within more tolerable limits. In other words, while findings from the study in 1994 indicated an improvement in inter-rater reliability after training, this study provided evidence of an enhancement in intra-rater reliability. It is, however, interesting to note that there was no clear distinction that could be drawn between inexperienced and experienced groups in terms of rater severity as the statistics showed fluctuation in both groups before and after the training. This result is confirmed by Davis's (2016) study which showed that while rater reliability improved with training, there was little effect on rater severity. This may demonstrate a need for further qualitative studies which can better our understanding of rater severity.

While Vaughan (1992), Weigle (1994, 1998) and Barrett (2001) examined rater characteristics in the short term after receiving training, Lumley and McNamara (1995) considered the effectiveness of training sessions over a longer period of time. One of their major findings was that the results of training did not endure, since large differences in rater severity were observed between time 2 (the time when the training was carried out) and time 3 (the time when the test administration was operated). Similarly, additional training was not found to have an impact on experienced raters in terms of inter-rater variability, rater bias (O'Sullivan & Rignall, 2007) or within-rater consistency (Knoch, 2011). The findings shed light on the need for standardising and certifying of rater judgement over a certain period of time.

Instead of examining the effect of rater training on a group of raters, Knoch (2010) investigated the impact of individual feedback as part of rater training on rater behaviours. The author was able to show that there was no relationship between raters' perceptions of feedback and the success of incorporating such feedback in their ratings although they generally felt positive about the usefulness of the feedback. Knoch (ibid), in her further discussion, argued that raters may not have more processing capacity available at the time of rating to return to the feedback and incorporate it in their subsequent ratings. To some extent, this seems to confirm the hypothesis proposed by Eckes (2008) about the influence of other factors rather than training on rater behaviours. Knoch et al. (2021) suggested that one way forward might be to identify rater types, as developed by Eckes (2012) and tailor feedback to match the decision-making process or personalities of the raters. Thus, it is important to understand individual raters' decision-making processes and the factors which can have an impact on the process, so that more detailed and personalised feedback can be provided.

In L2 speaking assessment, Davis (2012, 2016) investigated the effect of training on raters with different levels of rating proficiency when they performed their rating of TOEFL iBT speaking responses. The results revealed that the training influenced the way raters managed their scoring processes, in terms of more explicit attention paid, and fewer disorganised or unclear comments made. However, differences were found in the frequency of using and reviewing the exemplars, and in language features mentioned during rating. The raters also differed in their style of commenting, including the selection of topics covered and the amount of detailed explanation of specific points. Davis's (2012, 2016) research is comprehensive since the author used a mixed-method research design to further the understanding of the influence of two rater background factors (i.e., rater experience interacting with training) on their rating decisions. However, the findings mainly focused on accuracy of raters' interpreting the rating scales (H. J. Kim, 2015), and the conscious attention raters paid to specific language features in their rating processes (Davis, 2012), leaving other important aspects, such as the sequence of activities occurring in the mental rating process and/or the decision-making strategies raters used to arrive at their scores, not thoroughly attended to.

2.5 Raters' rating experience

Rater experience is another major factor that may affect the reliability of scores. Studies that investigate this characteristic have also led to mixed results. One of the early studies investigating the effect of expertise on rating behaviours was conducted by Cumming (1990). The most obvious finding to emerge from the verbal analysis of 13 raters was that while experts tended to gather information to make judgement on the quality of language, most of the inexperienced raters were more attracted to error identifying and editing. Additionally, the study appears to be one of the first studies that offered a detailed picture of the

differences in behaviours between novice and experienced raters by providing a list of behaviours and the frequency of those behaviours in the two groups. Similarly, Barkaoui (2010) tried to examine the effects of marking methods and rater experience on essay test scores and rater performance. The findings of Barkaoui's study lend support to Cumming's (1990) results that substantial differences between experienced and inexperienced raters do exist. For example, novice raters gave more emphasis to argumentation while expert raters put more emphasis on accuracy. In similar studies, 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).

A more detailed picture of the rating behaviour of expert and novice raters emerges in studies that have looked at more longitudinal data. Lim (2011), for example, was able to show that novice raters moved closer to the group average in terms of leniency and harshness soon after starting operational rating and that the same patterns were also observed for consistency. However, there was evidence in Lim's (2011) study showing that some novice raters may not be able to show rating consistency after a few months of rating. Steady increases in rater agreement with agreed scores were also observed in studies by (S. Shaw, 2002), although differences in rater severity persisted.

As a further attempt to investigate the differences among raters with varying degrees of experience, Isaacs and Thompson (2013) examined the effects of rater experience on their assessment of L2 pronunciation. Isaacs and Thompson were able to show evidence that raters diverged cognitively depending on their levels of rating experience. Raters with more experience were found to pay more specific attention to pronunciation errors. This rating behaviour was revealed through detailed characterisation and/or imitation/correction of TTs' speech. Moreover, the think-aloud and interview comments made by experienced raters were longer than those made by novice raters. Another difference between experienced raters and novices was the use of TESOL related vocabulary to describe L2 speech. Raters with more experience had more flexibility in applying professional knowledge in their L2 pronunciation assessment. Although the study revealed some cognitive differences between experienced raters and novices, it did not offer an adequate explanation for the differences. For example, the issue of novice raters' lack of TESOL vocabulary command was not verified. One possibility might be that the dimensions of the speech were heeded differently by novice raters due to a difference in perceptions and interpretation of the rating scale, rather than the inadequate access to vocabulary (Han, 2016). Moreover, it is essential in the context of assessing speaking as raters' fatigue can be one factor affecting their rating quality (Ling, Mollaun, & Xi, 2014); thus, how experienced raters can overcome fatigue to ensure their rating quality is of interest. Therefore, detailed understanding of the factors that might have affected raters' decisions of scores in their rating process is of significance (Han, 2016).

Another recent study (Lamprianou, Tsagari, & Kyriakou, 2020) investigating stability of rater behaviour over the course of 12 years (2002-2014) in the context of writing was able to provide a clear definition of how experience was operationalised in their study. Three different measures of experience were

used: (a) cumulative experience in the same exam, (b) cumulative experience in different exams, and (c) recent experience. The study found that the first two measures did not have significant impact on rating characteristics. Experience accumulated in other language exams was not transferable in terms of rating in this particular exam. This finding is concurrent with the findings of Huhta, Alanen, Tarnanen, Martin, and Hirvelä (2014) who discovered that previous experience of working with different types of rating scales does not necessarily enable raters to be ready to work with a new different rating scale. As the study employed quantitative methods, the researchers called for qualitative studies to further understanding of the issue.

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 how scoring behaviours might evolve, what scoring behaviours are related to experienced raters, and what scoring behaviours are associated with novice and developing raters remain obscure. Insights into these developments are of great benefit as they can inform rater training programmes and rater certifying processes. I would suggest that more qualitative studies are necessary to unpack this area of uncertainty. Thus, the third research question of this research project was formed in light of the issues discussed in this section.

2.6 Rater cognition

Rater cognition research is concerned with raters' mental processes in scoring TTs' performances. In the rating process, raters interact with three texts: the prompt, the essay/speech, and the rating scale. The rating scale plays an important role in second language speaking assessment because the scale content is closely related to the test construct (Fulcher, 2003), thus specifying

what raters should attend to, and ultimately influencing the validity of score interpretation and the fairness of decisions that educators 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 test validation.

Since rating scales are of great significance in language testing, the questions of how examiners assign a rating to a performance, what aspects of the performance they favour, whether experienced or novice examiners rate differently, and to what extent rating scales cause variation in evaluating written and spoken performance have drawn considerable attention from researchers. Recently, an extensive literature has grown up around this aspect of the rating context, trying to find sound answers for these questions; however, mixed results have been found in both written (Ballard, 2017; Barkaoui, 2007, 2010, 2011; Cumming, Kantor, & Powers, 2002; J. Huang, 2008; Lumley, 2002; Shirazi, 2012; Vaughan, 1992; Weigle, 1999) and spoken ratings (A. Brown, 2000, 2006; A. Brown, Iwashita, & McNamara, 2005; Orr, 2002; Pollitt & Murray, 1996; Yan, 2014). Although the focus of this section is based on reviewing studies on rater cognition in L2 speaking assessment, results in writing assessment research are occasionally mentioned since they may inform research into spoken rater cognition because of considerable similarities in the way raters interact with the rating scales.

Most of the research in the speaking assessment literature has focused on exploring features that raters pay attention to. Some differences were found in the way raters applied and interpreted the contents of the scale. A. Brown (2000), who investigated the rating process of 8 IELTS expert examiners found that many of the raters' comments consisted of inferences based on the TTs'

behaviour, and that these inferences often differed from rater to rater. Likewise, the verbal reports of 32 official FCE examiners in Orr's (2002) study revealed that raters heeded many aspects of the performance that are not relevant to the assessment criteria; for instance, some raters' comments referred to TT's age, gender, and TT's presentation of her/himself. Although L. May (2006) conducted her research in the context of paired candidate 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. L. May (2006) found that raters had a tendency for 'fleshing out' the criteria in the rating scale with features that were not explicitly mentioned. She found that as many as 30% of rater comments related to non-criterion-related aspects. These aspects included features such as the first impression of the candidate, the confidence of test takers as well as the complexity and logic of the candidates' ideas. In contrast, the findings of A. Brown's (2006) study showed remarkably few instances of examiners referring to aspects of performance not included in the scales. 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 I 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. In other words, 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 H. J. Kim (2015) who concluded that it is important to consider rater characteristics collectively in

studies investigating rating behaviour and in training raters. One of the findings of H. J. Kim's (2015) study which I deemed 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.

Furthermore, there have been no cognitive processing models developed which 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 (Knoch et al., 2021) 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 the study provided a detailed description and model of the rating process that raters followed. The raters in his study appeared to interpret the scale categories and descriptors similarly, but it was unclear how the raters connected the scale contents with the text quality in their rating processes. However, I argue that 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. I argue that the stages that speaking raters experience in their rating may be different from those which are proposed in Lumley's (2005) study for writing raters.

Looking at different rater decision-making behaviours in L2 assessment, Cumming et al. (2002) and B. A. 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 as they may suggest content and practice in rater training programmes, creating a clearer picture for raters to develop themselves. B. A. Baker (2012) attempted to classify writing examiners' behaviours based on the concept of decision-making style. The analysis of six raters' write-aloud protocols revealed that certain elements of the texts themselves can influence raters' scoring decision. For example, in one text, with some key information missing, three raters failed the TT, whereas the others awarded a borderline score. B. A. Baker (ibid) 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.

Another line of rater cognition research has focused on exploring the ease of using rating scales, band differentiation, and the effect of criteria order. A. Brown (2006) explored IELTS speaking expert examiners' (although no definition of expert was provided) interpretations and applications of the revised band descriptors and revealed an unclear distinction between adjacent levels and an overlap between scales. Such findings were also confirmed by a worldwide

survey of IELTS examiners' views and experiences (A. Brown & Taylor, 2006), where two problems with the revised band descriptors were noted. One of these problems was that the terminology of the descriptors was "subjective, vague or otherwise problematic to interpret", such as "sufficient, limited, basic, effective and occasional, wide range, etc.". The other was 'the difficulty of distinguishing particular adjacent bands' (A. Brown & Taylor, 2006, p. 16).

Understanding of how rating scale variation influences raters and rating processes has been extended through two important studies by Winke and Lim (2015) and Ballard (2017). The order of criteria in rating scales has recently been considered to affect writing examiners' rating behaviours. Winke and Lim (2015) found that the position of analytical rating categories potentially affected raters' mental formation of the rating scale, and consequently influenced raters' decision-making behaviours. By way of illustration, the first criteria on the left received most attention from the raters. The authors speculated that the observed scoring behaviours were due to ordering effects, or more specifically, primacy effects. This speculation was confirmed by Ballard (2017), even though the participants in her study were different from those in Winke and Lim's (2015) study. While all participants in Winke and Lim's (2015) study had ESL teaching experience and rating experience with similar types of ESL writing tests, all of Ballard's participants were raters without experience of teaching ESL and rating ESL essays. The participants in both of the studies received training on rating scales and rating processes. However, it is premature to speculate that the position of a criterion can influence the way raters weigh the importance of the criteria when scoring an essay. There might be the possibility that the raters in the two studies had internalised the rating scale; it was not because they were using their intuition to evaluate the essay. This points a need for research furthering understanding of these issues.

The varying results found in the studies conducted so far regarding rater cognition can be put down to different methodologies. Think-aloud protocols, write-aloud protocols, stimulated recalls and eye-tracking movements are common methods used to explore rater cognition, beyond interviews and/or questionnaires. In this section, the strengths and weaknesses of each method will be discussed.

Think-aloud protocols (TAPs) are a long-established method used in psychological and social research and have been widely used in rater cognition literature (used in Cumming et al., 2002; A. Brown, 2006). The method requires participants to speak out loud their thoughts while performing a specific task. There are two main variations in the way TAPs are used: concurrently and retrospectively (Suto, 2012); but according to Kuusela and Paul (2000), concurrent think aloud imposes less cognitive strain on participants' memory, thus being more effective in providing richer information about the thinking process than retrospective think aloud. However, one of the most common criticisms this method has received is that it tends to provide unnatural evidence of rating behaviour (Ballard, 2017). In other words, the verbalisation itself 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 than they do in their regular practice.

Another research method used is write aloud protocols (used in B. A. Baker, 2012), in which raters provide written comments on the reasons for their score decisions during their rating processes. B. A. Baker (2012) argued that the write aloud protocols had benefits over think-aloud protocols and provided rich data related to individual differences in rating behaviour although no further explanation was provided. However, B. A. Baker admitted that similar to think

aloud protocols this method has one side effect: it may change the nature of the rating processes, and the effect may be even stronger since the raters can have more time reviewing the notes, adjusting their comments as they go along.

Additionally, stimulated recall (SR – used in Winke, Gass, & Myford, 2011) is an introspective method resembling retrospective think aloud but relying less on memory. Participants' behaviours are video recorded and then replayed to help the participants recall their concurrent cognitive activity during that behaviour. Thereafter, a set of open-ended questions is asked during or soon after the video viewing. One merit of this method is that it does not alter the nature of the setting but can still reveal the thinking process when the participants talk (Suto, 2012). However, there are concerns related to the issue that participants can have chance to reorder their thinking processes in their verbal reports (Lyle, 2003) and that "some raters' comments gathered through SR appeared at times to go beyond their thought processes at the time of rating" (Winke et al., 2011, p. 52). Moreover, in response to the video viewing and the set of questions elicited by the researcher, the raters simply practice reflecting what they have watched, rather than recalling the original event (Gass, 2001). Gass also pointed out that researcher questioning skills can also alter the quality of the reflection.

Recently, eye-tracking (used in Winke & Lim, 2015; Ballard, 2017) has been given much attention by researchers in the field. One benefit of the eye tracking method is that it is unobtrusive, and therefore less likely to alter the natural setting. However, using the eye-tracking method could create some potential problems. Even though the possibility of changing the natural thinking processes is low, it cannot reflect the multi-dimensional cognitive process of human raters. For example, two of 11 raters in Winke and Lim's (2015) study were excluded because one had a vision problem (making the eye tracking data too unreliable) and the other had an extremely long rating duration (more than four times that

of any other raters). Furthermore, eye-tracking movement cannot be completed without a computer screen, at which raters are required to look. I would hypothesise that looking at computer screen and at a paper-based rating scale would reveal some differences in rating behaviours since raters are more familiar with scoring a performance while physically interacting with a rating scale.

This section begins by providing synthesis and evaluation of studies picturing raters' mental processes and arguing that there have been no cognitive-processing models of the rating process in L2 speaking assessment. It goes on to suggest that the rating scale itself, including content of descriptors and/or order of criteria might be the cause of raters' variations. Together with the issues discussed in the previous sections (2.4 and 2.5), the construction of research question 1 was made. This section also critically reviews different research methods used in the literature and suggests that among all the popular methods, data through TAPs can best reveal the nature of raters' thinking processes: where raters talk out loud during their ratings and rely less on memory, without causing cognitive strain on raters' minds nor changing the sequence of their thoughts (Ericsson & Simon, 1993). Thus, regarding my research project, which aims at furthering understanding of the mental processes of rating speaking performances, and the difference between novice and experienced raters, TAPs seem to be the most suitable way of collecting data in this respect.

2.7 Community of practice and sense of ownership

A community of practice (CoP) is depicted by Knoch et al. (2021) as an environment where raters can develop a sense of a group, a common goal and an opportunity for conversations and discussions. This sense of a community of practice may also affect rating behaviour. Similarly, Wenger, McDermott, and

Snyder (2002) considered a CoP as a "group of people who share a concern, a set of problems, or a passion about a topic, and who deepen their knowledge and expertise in the area by interacting on an ongoing basis" (p. 7). In educational assessment, Baird, Greatorex, and Bell (2004) argued that because marking criteria are relatively abstract and standards do not reside in the assessment criteria, the knowledge of how to apply mark schemes in practice is negotiated by a group of individuals and is based on tacit knowledge. Being involved in coordinators' meetings where changes to the rating criteria are negotiated also helps with the sense of being part of a team. Baird et al. (2004), in the context of GCSE English and History examiners in the UK, attempted to test this theory empirically and operationalised the features inherent in a community of practice as having access to benchmark samples or as having discussions between raters (as operationalised in three conditions: no coordination meeting, hierarchical coordination meeting and consensual coordination meeting). They found that none of these procedures showed an improvement in rating quality. However, questionnaire responses showed that the raters valued aspects of the community of practice they were engaged in, including coordination meetings where they were able to discuss harder-to-mark candidate responses. The study seemed not have been able to tap into the type of aspects which make such a community work and take ownership of the construct embodied within the assessment criteria.

Despite the initial research findings of Baird et al. (2004), it is now well established that CoP promotes individual raters to develop a shared understanding of the rating criteria and norms, thus leading to enhanced rating quality (Lamprianou et al., 2020). Conversations with colleagues about assessment standards and marking are found to be effective in gaining a mutual understanding of assessment criteria and increasing both tutor consistency and student satisfaction with feedback in an institution in Australia (Willey &

Gardner, 2011). In the context of accounting education assessment, Herbert, Joyce, and Hassall (2014) highlighted that learning in a CoP occurs in a situated setting where learners dynamically engage with information sources and social practices to develop themselves. Through this engagement, learning in a CoP is different from classroom learning which is often grounded in static and explicit knowledge. These ideas fit well with the notion that beginning raters develop their rating practice through their interaction with more experienced raters within the marking community. Similarly, as learning is a developmental process, expert raters can also learn from the diversity of rater expertise to refine and reshape their knowledge and practice. The findings show evidence of both aspects of a CoP, with newcomers gradually achieving acquisition of knowledge and skills in the community and experienced members improving their own learning through interaction with their peers (Herbert et al., 2014).

In the field of language assessment, little research has been conducted to understand how a CoP is operationalised and its impact on rating quality (Al-Maamari, 2016; Lamprianou et al., 2020). Lamprianou et al.'s (2020) study, the first study in the field of language assessment, was able to investigate quantitatively the concept of the CoP. The research team was able to show that regardless of prior experience, newcomers are more likely to be categorised as misfitting in a community with a large proportion of existing members. Additionally, when the community experiences a radical change in its membership, the most experienced raters can be classified as misfitting. The results indicate that a CoP plays a significant role in establishing standards and in influencing the rating behaviour of its members. However, the authors suggested that a qualitative study could have investigated how the stability of the rating group can be governed by micro-mechanisms. Insights from a qualitative study could have extended understanding of the way standards are negotiated by group dynamics and how these are renegotiated when the

community changes dramatically in terms of its membership. This knowledge could contribute to the development of rater training, including training tools and constructive feedback to raters.

2.8 Perceptions of rating criteria

2.8.1 Raters' perceptions of assessing fluency

Oral fluency is viewed as an important characteristic of second language speech, which explains why it is often the object of evaluation in testing second language skills. However, there is an ongoing lack of consensus regarding how fluency is defined and measured. In this sub-section, some interpretations of fluency are presented and a brief discussion of how speaking test descriptors interpret fluency are referred to. Finally, the studies which are representative of different research strands attempting to measure fluency will be critically discussed.

There are a number of definitions of the term *fluency*. Fillmore (1979) was one of the first researchers who characterised fluency in four dimensions, namely the number and duration of pauses, the quality of speech, knowledge of how to perform properly in different contexts, and creativity and wit in language use. While Fillmore's conceptualisation seems to consider fluency as an overall competence, Lennon (1990) distinguished EFL fluency in two senses: the broad sense and the narrow sense. Fluency can be broadly understood as "a cover term for oral proficiency" that corresponds to "the highest point on a scale that measures spoken command of a foreign language" (Lennon, 1990, p. 389). The narrow definition, on the other hand, refers to the temporal aspect of speech - one aspect of linguistic proficiency - and focuses on the ability to speak like an English L1 speaker (Lennon, 1990, p. 390). In a more recent study, Segalowitz (2010) distinguished between three facets of L2 fluency, namely cognitive fluency – "the efficiency of operation of the underlying processes responsible for

the production of utterances"; utterance fluency – "the features of utterances that reflect the speakers' cognitive fluency", which can be acoustically measured; and perceived fluency – "the inferences listeners make about speakers' cognitive fluency based on their perceptions of their utterance fluency" (p. 165). This interpretation takes into account both the temporal aspect of speech and the effects that speakers leave on listeners.

In language testing practice, definitions of fluency often include references to flow or smoothness, rate of speech, absence of excessive pausing, absence of disturbing hesitation markers, length of utterances, and connectedness. These characterisations are complex because they are not simply descriptions of a speaker's speech but also of a listener's perception of it. To illustrate this, in IELTS speaking band descriptors, the ability to "speak at length", and the number of instances of "repetition and self-correction", and "hesitation" are key terms to assess test takers' fluency. The descriptors also highlight the possible reasons for such pauses or hesitations, for example "content-related rather than to find words or grammar" in Band 9. However, according to A. Brown's (2006) study, fluency and coherence in the IELTS scale appeared to be the most complex since the descriptors for this scale covered many aspects of oral performance such as hesitation, topic development, length of turn, and use of discourse markers, which caused difficulty for examiners in making decisions. Often, examiners in A. Brown's study were unsure whether language or content was the cause of disfluency. Similarly, the description of fluency in the VSTEP speaking rating scale refers to hesitations and length and speed of delivery, and sources of such fluency features. For instance, the descriptors in Band 6 mention "hesitation may occur for grammatical and lexical planning". Considering these varying definitions of fluency and the way fluency is described in different rating scales in high-stakes test, it is necessary to define fluency for the present study. In this

study fluency refers to utterance (temporal aspect) and perceived (listeners' inferences) facets.

Different ways of interpreting fluency may lead to different measures of fluency. Several ways of measuring fluency have been proposed. One strand of research has explored similarities and differences of fluency characteristics between English L1 and non-English L1 speech (Bosker et al., 2014; Mulder & Hulstijn, 2011). Mulder and Hulstijn (2011) found that age, level of education and profession contributed to differences in English L1 speakers' lexical skills and speaking proficiency. The results of Bosker et al.'s (2014) study additionally revealed that disfluency variations in English L1 speakers shaped listeners' perceptions of English L1 speakers' fluency level. In other words, no consensus on measuring the fluency standards of English L1 speakers has been reached. As a result, the traditional assessment of non-English L1 speakers' fluency based on English L1 fluency standard should be reconsidered (Bosker et al., 2014).

Moreover, a growing number of studies have been concerned with assessments of fluency in the narrow sense (temporal measures) both with human raters and/or with the help of technology. The findings of these studies (Derwing, Rossiter, Munro, & Thomson, 2004; Kormos & Dénes, 2004) have cast doubt on accepted descriptors in fluency assessment. Particularly, while repetition and self-correction are two of the key terms in rating descriptors, Derwing et al. (2004) discovered that self-repetition appeared to be least related to other temporal measures. Different reasons for speakers repeating themselves may result in different listeners' inferences (Fulcher, 1993; Guillot, 1999). In addition, the study by Kormos and Dénes (2004) provided evidence to reveal that variables including pace, the mean length of runs and pauses, speech rate and phonation-time ration seemed to be important factors in fluency judgments.

Nevertheless, a high number of hesitations did not seem to be correlated with proficiency or fluency.

Furthermore, research in the literature provides evidence about the importance of the location of pauses such as mid-clause or end-clause positions (de Jong, 2016; Skehan & Foster, 2012; Tavakoli, 2011) and the character of pauses; that is filled pause or unfilled (silent) pause (Clark & Tree, 2002). The position of pauses is argued to distinguish L1 speakers from L2 speakers, rather than how much they pause. It is explained that L1 speakers pause at the end of clauses to prepare themselves with preverbal messages while L2 speakers are found to pause in the middle of clauses for lexical or morphosyntactic searching to convey the message (Skehan & Shum, 2017). Tavakoli, Nakatsuhara and Hunter (2020) also added that the distinction between lower (A2 and B1) level speakers and higher (B2 and C1) level speakers was the frequency of mid-clause silent pauses. They discovered that a higher proficiency level speech is characterised by fewer mid-clause silences. In terms of the character of pauses, Clark and Tree (2002) argued that both filled and unfilled pauses are indicative of language processing demands. On the other hand, filled pauses are found to facilitate communication as they can be used to draw attention to a particular point of speech, or idea organisation and communication strategies. However, the issue of whether raters are aware of and able to attend to these features while rating speech performances in high stakes testing contexts remains unclear.

The third strand of research into fluency assessment is to compare L2 learners' self-assessment of their fluency with English L1 speakers' assessment of fluency (Préfontaine, 2013). Préfontaine (ibid) discovered that French L2 learners' self-perceptions of fluency were moderately correlated with L1 French listeners' fluency ratings. Nonetheless, L2 learners may rate the speech of fellow L2 learners more harshly than English L1 speakers do (Rossiter, 2009). As a

consequence, two concerns should be taken into account: one is the question of whether L2 raters' perception is similar to English L1 speakers' perceptions, the other is whether L2 raters adopt English L1-like standards in their ratings. Similar to Davison's (2004) argument in the previous section, Préfontaine (2013) also called for more attention to be paid to the role of cultural norms in shaping the perception of fluency features since researchers will be able to construct richer descriptions of what L2 fluency entails by focusing on how speech performance is influenced by perceptions.

It can be seen from the reviews of aforementioned studies that definition and measurement of fluency as a criterion in speaking assessment is multi-layered and cannot be easily formulated. Moreover, researchers mainly use quantitative methods to measure temporal aspects of fluency and to correlate ratings of different groups of raters in their evaluation of fluency performance. Additionally, one gap in the literature review which can be perceived is that little research into the role of EFL teaching experience and language assessment knowledge in shaping raters' perceptions of fluency has been published. I would argue that findings from such studies will contribute to reshaping rater training programme contents and help raters reach closer agreement in assessing fluency.

2.8.2 Raters' perceptions of assessing pronunciation

The earliest test of pronunciation, which is known as the Biblical Shibboleth test, involved distinguishing one group of people from others. It was used in many societies as a password or a simple way of self-identification with fatal consequences if the 'wrong' pronunciation was produced (McNamara & Roever, 2006). This test raised concerns about the standards of speech against which L2 learners are judged in modern high-stakes language proficiency tests. However, the assessment of L2 pronunciation has been left behind since communicatively-

oriented theoretical frameworks fail to clearly and sufficiently address pronunciation competence (Isaacs, 2014). This section reviews and evaluates definitions of pronunciation, including the concepts of intelligibility in language testing and how it has been measured in different contexts. Moreover, discussion of the possibility of developing local norms for testing will be presented, along the lines of pronunciation assessment.

Pronunciation or, more broadly, the sound of speech, can refer to many features of the speech stream, such as individual sounds, pitch, volume, speed, pausing, stress and intonation (Luoma, 2004). Moreover, the issue of which principle, the 'nativeness principle' or the 'intelligibility principle' (Levis, 2006), L2 leaners should follow has sparked heated debate for a long time. The second principle refers to the ultimate aim of L2 pronunciation – to be intelligible to listeners. Levis (2006) attempted to distinguish between broad and narrow definitions of intelligibility. In its broad meaning, 'intelligibility' refers to listeners' ability to understand L2 speech (Levis, 2006, p. 252). In its narrow sense, 'intelligibility' is defined as the amount of speech that listeners are able to understand. How intelligibility is described and interpreted in high-stakes tests is discussed later in this section.

The first principle refers to the use of particular pronunciation norms – here it is English L1 standard as the final goal of L2 learners by reducing L1 traces from their speech. However, all languages have different regional varieties and often different regional standards; thus, these standards are valued in different ways. Moreover, Groves and Chan (2010) argued that it is unrealistic to expect English L2 learners "to conform to [English L1] norms in all respects, especially considering their input is largely from other locals, whose English exhibits the same features" (p. 48-49). In addition, the negotiation of the local and global dimension of the English language is a challenging task to address in

international examinations (Sewell, 2013) and in local testing contexts (Dimova, 2017; Harding, 2017). Generally, norms and standards have often been viewed and researched by international language testing bodies who deliver their tests in various local contexts (B. H. Huang, 2013; Winke, Gass, & Myford, 2013; Xi & Mollaun, 2011) and from second language acquisition to inform English teaching and learning (Bøhn & Hansen, 2017; Pinget, Bosker, Quené, & de Jong, 2014; Wicaksono, 2020).

The practice of assessing spoken English in a local context by its local people is a less researched topic. There are two recent important studies on L2 spoken English tests that sought to address issues in locally developed tests. The first study (Sewell, 2013) evaluated the degree of alignment between rater comments on their local TTs' speech and the criteria for international intelligibility in the Hong Kong context. This study revealed that there was alignment between rater comments and the criteria for international intelligibility. Sewell (ibid) provided a possible explanation for this. The raters might have been more likely to notice and comment on the features that affect intelligibility as these features stood out to them. However, its main limitation was the lack of detailed information about the scoring process and the actual effects of the comments on TTs' scores.

The second study (Nakatsuhara, Taylor, & Jaiyote, 2020) reports on how the rating descriptors of a locally developed test were used to assess Japanese leaners of English. The purpose of this study was to measure the quantity of L1-influenced (Japanese katakaba-like) words which tapped into the 'intelligibility' and 'L1 influence' aspects of the given rating scale. Three raters rated 23 video-recorded performances of Japanese L1 TTs and then watched videos of three TTs again, after which they discussed their reasons for the scores they awarded. Statistical analysis revealed that pronunciation was one of the two aspects of the

rating scale which was found to be easier than the other aspects. The raters reported one of their concerns while rating was that their familiarity with Japanese speakers' pronunciation of English enabled them to understand the TTs. This leads to the consideration of the EFL context in Japanese universities. As students in this context will be mainly interacting with their peers who are Japanese EFL learners and with their tutors who are familiar with typical pronunciation features of Japanese speakers of English, it is justifiable to be lenient about the impact of L1 influence on intelligibility and communication effectiveness. By doing so, the researchers concluded that the English variety spoken by Japanese speakers is included rather than excluded in the TEAP Speaking while the construct of the test and the usage of the test scores are still reflected. The context of this study shared considerable similarities to that of the current research project in the sense that the VSTEP is a locally developed test and is used to test Vietnamese L1 groups in Vietnam's context. However, the TEAP Speaking validation research project, particularly the part on the pronunciation scale, generated the data from TTs' outputs which were rated by three English L1 speaker teachers at Japanese universities whereas VSTEP speaking tests are rated by Vietnamese L1 raters teaching in Vietnam's higher education context. Moreover, little research has been done to unpack the issue of how local teacher raters are evaluating intelligibility, particularly in an EFL context, such as Vietnam, Korea, or Japan, where English is mainly used by people who share the same L1.

Another issue which is important to address in this section is how the concept of intelligibility is operationalised in rating scales in assessing speaking ability. The way pronunciation is modelled in existing rating scales has been criticised as being vague (Isaacs, 2014). Isaacs (ibid) argued that pronunciation descriptors do not often articulate a coherent construct. For example, in the public version of the IELTS speaking scale, the level four band descriptor reads, "uses a limited

range of pronunciation features; attempts to control features but lapses are frequent; mispronunciations are frequent and cause some difficulty for the listeners". The argument seemed to be supported by the findings in A. Brown's (2006) study as IELTS examiners in her study expressed a desire for more levels for pronunciation. They felt the scale did not distinguish TTs sufficiently and that fewer band levels meant the rating decision carried too much weight in the overall score.

Another important study which investigated the usability of pronunciation rating scale was conducted by Harding (2016). The study found that raters had difficulty in applying the CEFR phonological control scale to assess sample performances. The problems they encountered included clarity (e.g.: use of ambiguous terms such as natural), conciseness (lacking in detail), intuitiveness (e.g.: absence of self-repair) and theoretical currency (outdated view of English). These findings open up interesting implications for further research. With a specific focus on the development of pronunciation rating scales, Harding (ibid) suggested that it would be useful to conduct a study that integrates raters' suggestions into a revised scale with an improved set of descriptors.

This section reviews pronunciation-related key issues. I would argue that there is a pressing need to gain adequate understanding of the beliefs or the perceptions non-English L1 raters hold towards varieties of English, international intelligibility, and 'nature of speech' and discover if there are any differences in their ratings in these aspects.

2.8.3 Raters' perceptions of assessing lexical and syntactic complexity and accuracy

As lexical and syntactic complexity and accuracy are two common criteria assessed in a language speaking test (e.g.: in IELTS, TEAP, and VSTEP), it is

important to understand how these criteria are conceptualised and operationalised in high-stakes tests and to what extent this may contribute to raters' variety. This section, thus, discusses important studies in the field investigating these aspects.

Syntactic complexity is "the extent to which learners produce elaborated language" (Ellis & Barkhuizen, 2005, p. 139). Measuring the use of structures in terms of syntactic complexity is considered challenging. Length-based or subordination-based variables are often used to measure this aspect. T-unit (Bygate, 2001), C-unit (Mehnert, 1998) and the number of clauses per chosen unit (Foster & Skehan, 1996; Iwashita, McNamara, & Elder, 2001; Skehan & Foster, 1999) are typical examples of length-based variables. Examples of subordination-based variables include the number of subordinate clauses per clause (Wigglesworth, 1997) and the number of subordinate clauses per T-unit (Mehnert, 1998). However, Norris and Ortega (2009) suggest that these variables fail to capture the complexity of L2 learners' language features. In order to better reflect the multi-dimensional nature of L2 speech, they recommend using a number of variables, including:

- length-based variables (such as words per chosen unit),
- subordination-based variables,
- variables of phrasal complexity, and
- coordination-based variables.

The study (Norris & Ortega, 2009) reveals that coordination-based variables may be indicative of the beginner level while subordination-based reflects intermediate proficiency and variables of phrasal complexity advanced. However, the findings of Inoue's (2016) were not consistent with these suggestions due to a difference in contexts. While Norris and Ortega (2009) conducted their investigation of L2 writing, Inoue (2016) studied L2 speaking.

Their different findings suggest that definitions of syntactic complexity may be different according to the skill being investigated and that it is of necessity to clarify how raters perceive a complex construct as syntactic complexity in L2 speaking assessment. Thus, it is important to recall Halliday and Matthiessen's (2004) view of speaking in which they stated that in a number of respects that spoken language is different from written language. Speaking is usually characterised by less formal use of vocabulary, fewer full sentences as opposed to phrases, and speaking can contain repetitions, repairs and has more conjunctions instead of subordination (Halliday & Matthiessen, 2004).

Syntactic accuracy refers to the extent to which the target language is aligned with its rule system (Skehan, 1996). Examples of common variables for measuring syntactic accuracy include:

- the percentage of error-free clauses (Skehan & Foster, 1999)
- the percentage of error-free units (Robinson, 2007)
- the number of errors per unit (Bygate, 2001)
- and the number of errors per 100 words (Mehnert, 1998).

However, the issue of which variable is the most valid has not been agreed by researchers (Inoue, 2016). On the one hand, Bygate (2001) argued that it might be more suitable to measure syntactic accuracy by counting the number of errors per chosen unit instead of calculating error-free units. Bygate explained that this way of measuring does not complicate the actual occurrences of errors. On the other hand, Mehnert (1998) suggests counting errors per 100 words may be more useful for speakers with lower-proficiency levels because it can be problematic for this group of learners if the measure involves definitions of clauses and units. In response to these issues, Inoue (2016) sought to extend understanding of which variable is most valid in measuring accuracy. The differences of speaking performances were best captured through the errors per

100 words measure, as revealed by Inoue (ibid). However, one of the limitations of the study was that the majority of participants were at A2 or B1 levels, and more data at higher ends might have offered more concrete and reliable findings. The difficulty of evaluating the range of syntactic structures is reported by IELTS raters in a recent study by Inoue and her colleagues (2021).

In terms of **vocabulary,** it is a significant dimension of language use that is examined both from psycholinguistic perspectives (Skehan, 2018) and second language acquisition perspectives (de Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2012; Koizumi & In'nami, 2013). Quality of lexical resources is important, as generally agreed by researchers, since it determines the level of achievement that learners gain to fulfill their communicative functions (Crossley, Salsbury, & Mcnamara, 2015). Boers, Demecheleer, and Eyckmans (2004) also recognised the significance of the ability to use prefabricated multi-word lexical chunks such as fixed and semi-fixed expressions, collocations, pragmatic functions, idioms, etc. Read (2000) made an attempt to characterise vocabulary use by introducing a set of concepts to define the notion of lexical richness. The set consists of lexical density, lexical sophistication, lexical variation and number of errors in vocabulary use (Read, 2000). It is important to understand how these concepts are measured since they are often implied in rating scales in high-stakes tests. For example, the public version of IELTS rating scale reads as “has a wide enough vocabulary to discuss topics at length and make meaning clear in spite of inappropriacies” in Band 6 and “uses less common and idiomatic vocabulary skilfully, with occasional inaccuracies” in Band 8 of lexical resource.

Lexical density refers to the ratio of the number of lexical words to the total number of words in a text (Ure, 1971). However, different studies have different ways of specifying lexical words. Lu (2012), for example, considered nouns, adjectives, adverbs with an adjective base, and verbs (excluding model verbs,

auxiliary verbs, “be” and “have”) as lexical words whereas O'Loughlin (1995) identified all adverbs of time, manner and place as lexical adverbs. Moreover, Halliday (1985) and O’Loughlin (1995) characterised lexical density in spoken texts as having a lower density than written texts and being influenced by plannedness and degree of interactiveness. In addition, Uchihara and Clenton (2020) made an important point that the use of productive vocabulary measures is an under-researched area, compared with that of receptive vocabulary measures. The researchers argued that this might be due to the challenges of defining the construct of productive vocabulary knowledge.

Lexical variation, also labeled lexical diversity refers to the number of different words appeared in a text. Jarvis (2013) suggested that it consists of seven subdimensions, among which lexical variability is the most commonly operationalised construct of diversity. Lexical variability concerns how unique word forms are used differently in a text and is often measured by the type-token ratio (Treffers-Daller, Parslow, & Williams, 2018). The underlying assumption of lexical diversity is that the higher proficiency level language learners achieve, the wider variety of vocabulary items learners can use.

Lexical sophistication refers to “the proportion of relatively unusual or advanced words in the learner’s text” (Read, 2000, p. 203). Lexical sophistication has been traditionally operationalised in reference to lexical frequency-based databases (also known as corpora). Eguchi and Kyle (2020) listed several common measures, such as the lexical frequency profile (Laufer & Nation, 1995), and mean frequency scores (Graesser, McNamara, Louwerse, & Cai, 2004) and the Tool for the Automatic Analysis of Lexical Sophistication (Kyle, Crossley, & Berger, 2018). The working hypothesis of this approach is that higher frequency words are more easily acquired by learners while a high proportion of lower frequency words indicate a more elaboration of mental lexicon (Nation & Webb,

2011). The frequency-based operationalisation of lexical sophistication in earlier research has implications for task design as the tasks should be able to elicit low frequency vocabulary items (Inoue, 2016). Although the frequency-based approach has been widely applied for the past two decades, researchers have raised awareness of the need to re-examine the construct of lexical sophistication from different perspectives (Crossley, Salsbury, McNamara, & Jarvis, 2011; Kyle & Crossley, 2015).

One among very few studies exploring lexical sophistication in speaking context was conducted by Eguchi and Kyle (2020). The researchers suggested that lexical sophistication comprise of 5 dimensions as listed below:

a) rareness
b) conceptual features
c) formal distinctiveness
d) accessibility and
e) association strengths of the multiword units.

Multidimensionally operationalised lexical sophistication helps to identify supplementary characteristics of 'advanced' vocabulary used by language learners. This study contributes to better understanding of the construct by providing a nuanced description of lexical use. It also brings light to the need for further investigation into the ways human raters evaluate this construct in their rating process.

This section reviews studies investigating how different indices are suggested and validated to measure aspects of grammar and vocabulary in understanding L2 language proficiency. It can be seen these studies relied heavily on quantitative research methods and very few studies looked at the ways human raters attend to and assign scores for these assessment criteria. The issue of

whether raters' minds are able to attend to the suggested indices remains unclear. Therefore, I proposed the need to find answers to the issue of what cause disagreement among all the raters in their rating decisions of scores (the current study's research question 2) to better our understanding of why they behave differently from each other in their rating processes.

2.8.4 Raters' perceptions of assessing speech's discourse competence

Different aspects of language are acknowledged to contribute to overall language proficiency, which feeds into L2 speaking construct (Iwashita, Brown, McNamara, & O'Hagan, 2008). Among them, discourse competence is an important aspect which refers to the ability to produce and comprehend unified pieces of oral or written texts beyond sentence levels (L. Bachman & Palmer, 1996). To explore discourse competence, one way, suggested by J. Y. Kang (2005), is to consider the degree of cohesion and coherence demonstrated in performance.

Cohesion generally refers to the ability of using cohesive devices to facilitate comprehension of the text, "as opposed to a sequence of sentences that would not be considered a text" (J. Y. Kang, 2005, p. 264). While it appears to be easy to arrive at a unified definition of cohesion, it requires more attempts to define coherence. In written context, coherence involves a degree of unity between sentences within a text, which can be done through semantic relations (Cameron et al., 1995). In spoken texts, Seedhouse and Harris (2011) foreground logical links between sentences and ideas presented as an important aspect of coherence. Seedhouse and Harris also acknowledge the use of cohesive devices in order to be coherent in spoken texts. Celce-Murcia, Dörnyei, and Thurrell (1995) added one important yet challenging aspect of coherence, which is "the degree to which sentences or utterances in a discourse sequence are felt to be interrelated rather than unrelated" (p. 15). It is important to unpack the

inclusion of “felt to be” as it seems to indicate the puzzling aspect of coherence, particularly in spoken discourse. The degree of coherence tends to take into account the role of the listener in perceiving and understanding the unity of the text. Thus, Iwashita, May, and Moore (2017) argued that coherence can be considered co-constructed since it involves essential interaction between the content of the text and the knowledge of the listener. From these definitions, coherence can be characterised as “interrelated, unified and meaningful to the listener” (Iwashita et al., 2017, p. 9). To sum up, while cohesion refers to the properties within a text, coherence refers to its “contextual properties; that is the way in which it relates to and makes sense in the situation it occurs” (Paltridge, 2000, p. 139).

There are three important studies in the field which attempt to explore the notion of coherence in speaking contexts. The first study was conducted by Iwashita and Vasquez (2015) who used two measures to explore coherence in the context of an IELTS speaking test, including the identification of theme and rheme development and text generic structure. The findings showed that the number of discourse features (e.g: use of a wider range of conjunctions, more accurate use of referential expressions) and the sophistication level of patterns distinguished higher proficiency TTs from lower-level ones. The study (ibid) also examined features of discourse competence through the use of cohesive devices (e.g.: reference, ellipsis and substitution, lexical cohesion, conjunctions). Statistical analysis revealed that TTs with higher proficiency level demonstrated better control of cohesive devices, and their referential expression was more accurate than lower proficiency test-takers. The findings provide important implications for content of rater training in terms of raters’ attention to identifying cohesive devices used by test takers and logical links between sentences and ideas.

Another study which contributes to extending understanding of features of discourse competence was conducted by Iwashita et al. (2017) who compared features of discourse competence and vocabulary use across levels and tasks in the Aptis Speaking Test. The research team used both quantitative and qualitative methods to examine the role of cohesive devices, including conjunction, reference, lexical cohesion and vocabulary use. For coherence, a qualitative method was used to investigate the degree of relevance of TTs' responses to the questions asked, and the degree of textual unity. The study discovered that conjunctions are found across all levels of spoken performance. In terms of coherence, high-scoring performances were distinguished from low-scoring ones in a number of features. For example, TTs with higher scores tended to provide longer responses, be able to develop topics by using lexical chains, and connect ideas by using pronouns and conjunctions. Although the results have important implications for the design of rating scales and rater training, Iwashita et al. (2017) suggested that further study is needed to unpack how raters arrive at a score to gain clearer understanding of how the notion of cohesion and coherence are operationalised from the raters' perspectives.

A different approach to examine features of discourse competence was taken by Seedhouse and Harris (2011), who focused their attention on topic development and management in TTs' responses to the IELTS speaking test by using Conversation Analysis as the main data analysis method. Their findings identified features of high- and low-scoring performances in relation to topic development. For instance, TTs achieving higher scores were found to produce extended turns, elaborate topics with multiple examples, construct coherent answers with cohesive devices and use less common lexical items. An interesting aspect that emerged as relevant to TTs at higher end of the scale was that "candidates who achieved a very high score typically developed topics that constructed the identity of an intellectual and a (future) high-achiever on the

international stage" (p. 25). The results of the study provide several recommendations to IELTS rater training in terms of use of follow-up questions, the importance of examiners following their briefs, and of explicit marking of topic shift.

Another issue which is important to address in this section is how the concept of discourse competence is operationalised in rating scales in assessing speaking ability. For example, in the context of the IELTS speaking test, coherence is a joint criterion with fluency. Two relevant descriptors in Band 9 are: "speaks coherently with fully appropriate cohesive devices" and "develops topics fully and appropriately". Implicit in these descriptors is the ability to explicitly use cohesive devices and develop topics in a way that is relevant to the task. Another instance when coherence is operationalised in an international English language test is seen in the rating scale of TOEFL iBT for independent speaking tasks. The descriptor for a score of 4 states that the "response is sustained and sufficient to the task. It is generally well developed and coherent; relationships between ideas are clear". Thus, the key aspect of coherence in the context of TOEFL iBT is the semantic relations between ideas that are implied in the way TTs develop their ideas. From these examples, Iwashita et al. (2017) argued that there are different ways of operationalising the notion of discourse competence, and that the distinction between coherence and cohesion remains unclear.

These reviewed studies, from TTs' outputs, provide important insights into the understanding of the construct of discourse competence and how it is operationalised in international language tests by using qualitative methods. It can be seen that there is a need to investigate how these notions of cohesion, coherence and topic development are perceived and assessed by raters. The study can provide valuable information to further understanding of the construct.

2.9 Conclusion

In seeking to understand how different factors may influence raters' scoring decisions in speaking contexts, I have drawn on literature that discusses the potential impact of rater background, raters' rating experience, rater cognition, CoP, and raters' perception of assessing fluency, pronunciation, syntactic and lexical complexity and accuracy and discourse competence on raters' behaviours and score interpretation, as suggested in Knoch et al.'s (2021) model. The complex interaction between these variables in different contexts has resulted in contradicting findings. To date, research studies do not show agreement on the consequences that each factor facilitates. Thus, rating processes and raters' decisions can be considered as an unpredictable phenomenon and can be influenced by a number of factors. Hence, more research on rating processes and raters' decisions in speaking tests is required to further understanding of why raters behave the way they do. It is possible that there are still different interactions between factors influencing their rating processes and rating decisions in different contexts, which have not yet been described. For that reason, these issues are worthy of investigation in this research project and in the context of this study.

This study, therefore, examines the rating processes and the extent to which variables (e.g.: background, rating experience, cognition, perceptions, and contexts) influence their rating decisions. It aims to provide comprehensive and contextualised descriptions in order to present an authentic picture of the rating processes and rating practice development in the research context. In the case of EFL local teacher-raters evaluating their local learners of English, the study could yield insights into how the raters encounter the rating processes with their beliefs, experience, and practices.

Chapter 3: Methodology

3.1 Introduction

This chapter presents the research methodology adopted to answer the following research questions:

1. What are the mental processes of rating speaking performances?
    - What differences between experienced and novice raters, if any, are seen in attention paid to language features described in the rating scale?
    - What differences between experienced and novice raters, if any, are seen in decision-making strategies?
    - What differences between experienced and novice raters, if any, are seen in attention paid to test-taker proficiency levels?
2. What are the factors that cause disagreements among all the raters in making score decisions?
    - In what ways and to what extent do these factors affect raters' decision in their ratings?
3. In what ways do raters develop their rating practice?

The chapter starts with a discussion of the researcher's ontological and epistemological perspectives, which informed the choice of an interpretivist research paradigm and a qualitative approach to further the understanding of the inter-relationship between factors which affect raters' scoring behaviours and the ways their rating expertise develops. I then present the underpinnings of the phenomenological strategy used in this study: interpretative phenomenology. Following this I discuss the limitations of the chosen methodology and then position my choice of interpretive phenomenology over other qualitative strategies, critically evaluating those alternatives in relation to

the subject and research questions with which this study is concerned. The chapter also covers the sampling technique and the participants, as well as the data generation techniques including moderation, think-aloud protocols and semi-structured interviews. Moreover, the processes of piloting research instruments are described. The chapter also discusses the associated ethical considerations and issues of ensuring the trustworthiness of the study, such as credibility and dependability. Finally, my role as a researcher is also presented.

3.2 Justification of the paradigm adopted

It is important to unpack the two terms ontology and epistemology as they inform the philosophical stance establishing important assumptions about how researchers view the world and how their research is framed and bolstered (Cohen, Manion, & Morrison, 2007).

Ontology is a system of beliefs that reflects how an individual looks at and makes sense of what constitutes reality (Crotty, 1998). In other words, ontology is associated with the nature of existence, i.e., whether the structure of reality needs to be perceived as objective or subjective. Epistemology is concerned with what we can know about reality and how we can know it (Hammersley, 2013). In this part, ontology and epistemology are discussed together as one informs the other.

Those who view reality objectively often believe that: “Reality exists ‘out there’ and is driven by immutable natural laws and mechanisms. Knowledge of these entities, laws, and mechanisms is conventionally summarised in the form of time- and context-free generalisations” (Guba, 1990, p. 20). In this regard, the existence of objects and their meaning in the world are in advance of and independent of any consciousness of them. In this sense, these objects can be calculated and are thus verifiable. In other words, the world in the eyes of the

positivists is highly logical and well-organised. It is: “a world of regularities, constancies, uniformities, iron-clad laws, absolute principles” (Crotty, 1998, p. 28). Consequently, according to Crotty (ibid) the primary way to know these objects in the world of positivism is through value-free scientific methods and independent observations. This view has been popularly applied in studies in language testing and assessment since it has been considered as an effort to find ways of making tests reliable and valid (Spolsky, 2017). Spolsky (2017), a highly influential scholar in the field of language testing, stated that:

> *“Looking back over the half-century during which language assessment has developed into an identifiable academic field as well as a major industry, there are several trends which are worth identifying. One, particularly relevant to the academic field but with strong influence on practical test development, has been the effort to overcome what was recognised a hundred years ago as the unavoidable uncertainty of examinations (Edgeworth, 1888). Once statistical methods of establishing reliability were found, replacing single individual measures like essays with large numbers of objective items lending themselves to appropriate statistical treatment, testers could argue that their test was reliable […]” (p. 378)*

L. Bachman’s (2004) view seemed to be in parallel with Spolsky’s when arguing that psychometric and statistical methods play a vital role in language testing research since they help to provide an important kind of evidence to support test use. Moreover, McNamara and Roever (2006), who have published numerous articles and books in the area of language testing, also admitted that “fairness plays an important role in traditional psychometric work on testing, and a variety of [statistical] approaches have been developed to detect unfair items and investigate unwanted influences of test-taker background factors” (p. 127). In other words, to discover the characteristics of a test, a rigorous, objective and scientific method – practices of measurement – are usually needed. Therefore, tests and issues related to tests are often considered to be neutral, objective and value-free without the involvement of feelings and emotions (McNamara, 2009; Shohamy, 2001).

Another view of what constitutes reality is that reality is subjective, multiple and contextual, rather than existing 'out there' to be objectively found by researchers (Saldaña, 2011). These interpretivists not only embrace subjectivity but also value the context and culture which form reality. In this light, each individual has their own view of reality; they can make sense of and interpret a phenomenon differently, but their culture and context has contributed to shaping the reality. This is also the ontological position which has informed this study. I perceive that social phenomena and their meanings are personal and variable and as Bryman (2008) argued, that they are products of consequent actions and perceptions of those who perform the actions. This ontological position is in line with the nature of the research phenomena: scoring decisions and rating practice development. Scoring decisions and rating practice are the ultimate results of individual rater's perceptions of the VSTEP test, the test tasks, their own understanding of TTs, of the rating criteria they are applying, their personal accumulation of English knowledge and their scoring experience. All of these factors vary from each other; hence, as Hammersley (2013) argued, different interpretations of the world could lead to different reactions to the same situation. Scoring decisions and rating practice development are thus viewed as existing, however, not only as the participants' perceptions, beliefs and practices, but also as social constructions, i.e., the interaction between their perspectives, their thought and the language of the wider society.

Holding this ontological position, my epistemological philosophy, which adopts interpretivism, demonstrates that reality is based on social interaction, bound to contextual and cultural values; thus, the interpretation of reality should not be reduced to be simplistic, but rather gained through personal experiences arising from particular situations. Glesne (2016) argued that it is important to interact with people in their social contexts and talk with them about their perceptions. The aim of my study was to develop insights into the rating experience of VSTEP

raters from their subjective perspectives. Thus, the interpretivist approach enabled me to access the inner world of the participants as Giorgi and Giorgi (2008) suggested. It provided a rich understanding of how scoring decisions were made and how rating practice was developed. Turning back to positivism, positivists may argue that scoring patterns and rating practice can be revealed through measurements where raters' scores are collected and calculated. However, those numbers cannot tell the researcher in what ways the scores are formed, what raters mean by giving the scores and why differences in scoring patterns exist. Interpretivism, thus, can attempt to fill this knowledge gap by providing "important evidence just where scientific research was inadequate" and that "there is loss in the reduction of aspects of human behaviour to numbers" (Shipman, 1997, p. 38).

3.3 Research approach

Qualitative approach

Qualitative inquiry increases understanding of the everyday lives of particular people and the meaning of their actions. It identifies different sources of facts in the world such as people, actions, beliefs and interests and how those sources may form a difference of meaning (Erickson, 2018). Bearing in mind those purposes, qualitative research is often associated with research where data are not in the form of numbers (Babbie, 2015). Thus, a qualitative approach can "capture what actually takes place and what people actually say, in other words, perceived facts" (Patton, 2002, p. 28). Accordingly, I adopted the qualitative approach which enabled me to gain a deeper understanding of speaking raters' perceptions and beliefs, and how their scoring decisions were made. It also provided a rich understanding of the way the raters developed their scoring practice. Qualitative research was appropriate for my study because I was able to gain a deeper understanding of the phenomenon and to hear the raters'

voices of the factors that may affect their scoring decisions and affect their professional development. Therefore, qualitative research was adopted to understand the social world from the perspective of the raters, which quantitative studies cannot reveal.

3.4 Research design

Phenomenology

An important strand of thinking within interpretivism derives from the phenomenological movement in philosophy (Hammersley, 2013). This argued that the social phenomenon and its meaning are the results of immediate experience; thus, phenomenologists require careful description of the experience (Hammersley, 2013).

The descriptive tradition of phenomenology, which was established by Husserl (cited in Crotty, 1998), focuses on unpacking the lived experience of those being studied, without any interpretation by the researcher (Crotty, 1998). This school of phenomenology is influenced by Husserl's enthusiasm to ensure that all research is grounded in scientific framework, i.e., conducted through objective and true descriptions of the phenomenon by the researcher without any influence from them. In developing Husserl's work further, Heidegger, Merleau-Ponty and Sartre (cited in Crotty, 1998) emphasised the significance of the relationships between individuals and the world they live in (Crotty, 1998). This interpretive tradition of phenomenology attempts to understand how people make meanings out of their activity and their relationship to the world. Researcher presuppositions, including beliefs, prior assumptions and past knowledge of the relevant literature are seen here as an integral part of the research process, rather than being eliminated in the descriptive tradition (Lawthom & Tindall, 2011).

I subscribe to this epistemological position of interpretive phenomenology since it "maintains a central focus on the ways in which people make meaning of their experience, whilst being aware of the influences that broader social structures have on those meanings" (Lawthom & Tindall, 2011, p. 10) and acknowledges the primary role of the researcher in the research process (Smith, Jarman, & Osborn, 1999). To me, the lived experience of VSTEP raters from their own perspective is central, although there is acceptance that the participants can be in the same setting of VSTEP speaking tests but may experience this differently because of the influences that their personal structures and broader social structures have on them. These individual realities are socially and experientially constructed, are thus personal, multiple and delicate (Denzin & Lincoln, 2000; Guba & Lincoln, 1994). Moreover, interpretive phenomenologists argue for, and I agree with this, "their embeddedness in the world of language and social relationships, and the inescapable historicity of all understanding" (Finlay, 2009, p. 11). Therefore, the researcher's assumptions and presuppositions about the phenomena under investigation are inextricable from the research findings (Lopez & Willis, 2004).

This phenomenological approach was adopted to achieve the aims of my study. One aim of my study was to further an understanding of the scoring process of VSTEP raters so that I could understand what factors may affect their scoring decision in this process. In other words, I attempted to understand what meaning VSTEP raters made and how they made these meanings in relation to their scoring as they experienced the scoring process or their lived experiences of evaluating VSTEP speaking performances. In order to achieve this aim, phenomenology was suitable since it "seeks the psychological meanings that constitute the phenomenon through investigating and analysing lived examples of the phenomenon within the context of the participants' lives" (Giorgi & Giorgi, 2008, p. 28). Additionally, phenomenologists believe that behaviour is a

consequent product of our reaction to past experiences or, as Keen (1982) writes, “behaviour is an expression of being in the world” (p. 27). From this perspective, action and individual agency are considered as embedded in a broader social context (Leavy, 2014). Phenomenology helped me unpack the factors which were involved in the participants’ decisions: some of those may relate to their teaching experience or their perceptions of speaking assessment criteria which could be revealed through the participants’ memories of the experience and their actual experiences of the phenomenon. The voices from these raters then will be helpful for test administrators, policy makers and other stakeholders in Vietnam and other similar contexts.

Another aim of my study was to understand how the raters develop their scoring expertise. In this regard, I tried to “involve a return to experience in order to obtain comprehensive descriptions that provide the basis for a reflective structural analysis that portrays the essences of the experience” (Moustakas, 1994, p. 13). In this sense, this aim seemed to fit the concept Sartre (1996) presented in his approach to phenomenology because the participants’ professional development processes may depend on a number of factors, including the activities they have engaged in and “the embodied, interpersonal, affective and moral nature of those encounters” (Smith, Flowers, & Larkin, 2009, p. 21). Phenomenology allowed me as a researcher to access not only a rich description of the participants’ past experience of scoring speaking performances but also their reflection of how their scoring behaviours had changed and what had made the changes according to their perspective. Using phenomenology helped fill the gap in the literature by unpacking in detail the ways raters think, do, feel and reflect in their scoring procedures over time, something which quantitative data cannot reveal.

However, phenomenology does receive some critiques. First, the process of "reading between the lines" has generated uncertainty as the question of how far the interpretation can go beyond the actual quotes of the participant is of concern. I agree with Wertz (2005) when he argued that: "interpretation may be used, and may be called for, in order to contextually grasp parts within larger wholes, as long as it remains descriptively grounded" (p. 175). Thus, I provided a rich description of the individual rater's experience of scoring VSTEP speaking performances and of developing their expertise alongside my interpretation of the phenomenon so that I could maintain "the spirit of the phenomenological tradition that prizes individuality and creativity" (Langdridge, 2008, p. 1131). The second critique is researcher subjectivity. Some phenomenologists who follow descriptive phenomenology emphasise the reduction as a process of minimising the imposition of oneself on the data. However, since my study adopted interpretive phenomenology, I would deny that it is possible to set aside or bracket the researchers' experience and understandings. I agree with Finlay's (2008) argument that researchers need to bring "critical self-awareness of their own subjectivity, vested interests, predilections and assumptions and to be conscious of how these might impact on the research process and findings" (p. 17). I discuss this further in Section 3.10 *The role of the researcher*.

3.5 Sampling

Since understanding particular phenomena in particular contexts is a major concern of phenomenology, phenomenological studies require small sample groups (Smith, 2008; Smith et al., 2009). Moreover, participants are purposefully selected so that they can grant access to a particular perspective on the phenomena under study. However, within those groups, participants need to be as homogeneous as possible so that any differences found in people's experiences are down to different world of lived experience, rather than

different circumstances (Smith et al., 2009). The degree of specificity necessary in a sample depends on the focus of the study (Smith, 2008).

The boundaries of the sample in this research project were defined by the topic itself (Smith et al., 2009). In this study, I looked for raters who were teachers of English as a foreign language (EFL), had successfully attended VSTEP speaking rater training programmes and had performed their rating for at least a year so that they could offer insight into their scoring behaviours and scoring practice development. All the participant raters worked in the same institution. The criteria to select participants distinguished this study from other previous studies in that I 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. The reason for this was that the number of ratings was likely to be a more accurate identification of scoring experience since experienced teachers were not necessarily experienced VSTEP raters. Besides, experienced raters, based on this criterion, provided more insightful information about their scoring experience and their scoring practice development. More importantly, novice raters were often seen as those who have no previous experience of rating and/or no prior experience of EFL/ESL teaching (Ballard, 2017; Goh & Ang-Aw, 2018; Winke & Lim, 2015), which tends not to reflect the current situation in most testing contexts. Raters, particularly in high-stakes tests, are often required to obtain a certain amount of teaching experience and are certified through rater training programmes. In my study, novice raters were defined as EFL teachers who had occasionally performed ratings in one year, as compared with experienced ones who had regularly performed their ratings for more than one year. Hence, the participants could represent their own perspectives and/or their own lived scoring experience in their own context.

Participants

Based on my work experience I identified raters who met the criteria I set out above and contacted seventeen raters, fourteen of whom agreed to participate and two of whom did not because of a clash of schedule, while one did not reply. Table 3.1 below illustrates the rating experience and teaching experience of the participants.

Table 3. 1: Participants' information

	Raters	**Rating experience**	**EFL teaching experience**	**Number of VSTEP trainings attended**	**Educational qualification**
1	Daffodil	Experienced	over 10 years	Over 5	MA
2	Orchid	Experienced	over 10 years	Over 5	MA
3	Tulip	Experienced	over 10 years	Over 5	MA
4	Sunflower	Experienced	over 10 years	Over 5	MA
5	Lotus	Novice	over 10 years	Under 5	MA
6	Daisy	Novice	5-10 years	Over 5	PhD student
7	Lavender	Novice	5-10 years	Under 5	MA
8	Rose	Novice	over 10 years	Under 5	MA
9	Hyacinth	Novice	under 5 years	Under 5	MA
10	Jasmine	Novice	under 5 years	Under 5	MA
11	Lily	Novice	under 5 years	Under 5	MA
12	Iris	Novice	5-10 years	Under 5	MA
13	Peony	Novice	5-10 years	Under 5	MA

14	Lilac	Novice	Over 10 years	Under 5	PhD

The sample was fairly homogeneous as they all worked in the same institution, received the same training programme, and performed their scoring on the same test. The participants were purposefully selected so that they could provide insight into their scoring experience and the development of their rating practice. The differences, if found, then might come from their own world of experience which may include their own perceptions of scoring speaking performances and/or their teaching and scoring experience, but not from different circumstances. However, it is important to note that the researcher did not have profound knowledge of the participants' biography including their own experience of being tested, their attitudes to testing and their past experiences of education and language learning. Since the study tried to unpack their lived experience of being a VSTEP rater assessing spoken English, the lack of the knowledge on these biographical aspects of the cohort may have had an impact on the study. Sheehan and Munro (2017) argued that in classroom contexts, experiences of being assessed in previous education tend to play a role in shaping how teachers perceive and conduct their assessment activities later in their teaching career. Although the participant raters in this study performed their rating in a slightly different context – a high-stakes test, compared with the participants in Sheehan and Munro's (2017) study, it could be argued that their educational and testing experience could partly contribute to explaining why they behaved the way they did in their current job.

All the participants and the name of the institution were given pseudonyms. More details about pseudonyms are considered in the next section. All the participants were female except one who was male. However, because the gender of the participants was not a concern in this research project, she/her was used to refer to all of the participants.

The VSTEP speaking test

The speaking test consists of three parts (see Appendix 1 for a sample of a test). In Part 1 (Social interaction), the TTs are required to answer 3-6 questions on two different topics. This part lasts approximately 3-4 minutes. In Part 2 (Situation), the TTs are given a situation with three options to select. The TTs are required to select the best option and explain the reason(s) why they have chosen that option in preference to the other two. The TTs have 3-4 minutes to explain their choice. The third part of the test, which lasts 3-4 minutes, provides the TTs with a topic and a mind map of suggested ideas of how to develop the given topic. The TTs can use the suggested ideas or his/her own ideas. If time allows, the TTs discuss several follow-up questions upon completion of Part 3 (Topic development).

VSTEP speaking test responses

With relevant permission, I was able to gain access to 15 TTs' responses from several past VSTEP test administrations. The data set contained scripts of VSTEP speaking tests, test takers' scores and recordings. The responses are divided into 6 sub-categories, based on their scores in the speaking component, as shown in the following table (Table 3.2). The categories were based on the proficiency level of the Common European Framework of Reference (CEFR), which VSTEP tests are aligned to. The reasons for selecting these responses were (1) to see if there were differences in attention paid by raters to different proficiency levels and (2) to see if there were differences in raters' scoring decisions for the responses which were at the borderline, and what was involved in the decisions made by raters when giving the final scores for those borderline responses.

Table 3. 2: Selected VSTEP speaking responses

Level	Number of TTs
Border line between A2-B1	3
B1	2
Borderline between B1-B2	2
B2	3
Borderline between B2-C1	3
C1	2
Total	15

VSTEP rating scales

TT responses were scored using the 5 criteria listed in the ten-point rating scale. The criteria comprise Grammar, Vocabulary, Pronunciation, Fluency and Discourse management. The scale aims to examine test takers' proficiency levels at B1, B2 and C1 according to CEFR levels. An example of the rating scale is provided in Appendix 2.

3.6 Ethical considerations

The ethical issues of this study throughout the process of data collection and data analysis were considered under the umbrella of the British Educational Research Association's (BERA, 2018) guidance and International Language Testing Association Code of Ethics (ILTA, 2018). One of the underpinning principles in the guidelines is to assure that the study poses no harm to the

participants or the researcher and that it is operated within an ethical framework of respect. In order to do this, some ethical issues needed to be addressed.

First, I gained ethical approval from the School of Education and Professional Development to conduct my study. Thereafter, it was important to obtain consent to access the research site to legitimate the data generation process (Aldridge & Levine, 2001; Bell, 2014). I sought permission from the president of the institution and the director of the centre to access the data and cooperate with their staff in my data generation process. As I used to work in the centre, I understood the policy I needed to follow to keep the confidentiality of the relevant information as a top priority. Therefore, the materials including VSTEP speaking test scripts, VSTEP rating scales and 15 recorded performances were treated with strict confidentiality. The test scripts and rating scales were delivered to the raters for their use in an authorised area and collected after use from the raters by the researcher. The researcher kept the scripts and rating scales in a locked drawer in her office to make sure no one could access those materials. In order to keep the TTs' information confidential, the greeting part of each recorded performance was extracted since the test takers were required to speak out loud their full name and date of birth in the greeting part. The recorded performances were given pseudonyms of a number from 1 to 15 when given to the researcher and to the raters. Finally, all of the materials were discarded in the test security room immediately after the completion of the data collection procedure under the supervision of a member of the team.

Second, in terms of the participation invitation, I sent out an email with the information about my research project, including its aim and the main features of its design, the description of the tasks they were expected to do and their rights of participation in the form of an information sheet (see Appendix 4). I

also talked through all of the information with the participants and showed willingness to answer all of their questions when I met them for the first time before conducting any data collection activities to make sure that the decision they made was “based on adequate information about the project” (King, 2010, p. 100). They were also informed about having the moderation discussion, their verbal report and the interview recorded (more discussion of these methods in section 3.7). When they were happy to take part in the project, informed consent was sought. It is important that the participants were clearly informed of the purpose and the significance of the study as Cohen, Manion, and Morrison (2011) and Creswell (2013) reported that such action encourages participants to give comprehensive responses and consequently increases the reliability and validity of research findings. Consequently, all participants signed a consent form before the procedures of data generating started as recommended by Bryman (2008).

Moreover, it is important to inform the participants that they have the right to withdraw from the project at any point, without any requirement to explain their decision and without any subsequent consequences for them. Bearing this principle in mind, I kept a secure code sheet matching participant names to code numbers in the data set so that I could identify data with individuals, providing it if they later decide to withdraw their data. However, none of the participants indicated an intention to withdraw from the study.

Another ethical issue is confidential access to participants’ personal information disclosed in the course of the study. I gave each participant a pseudonym at the beginning of the data collection process to make sure that it did not reveal the identities of the participants. As the gender of the participants was not an issue of investigation of this study, she/her was used to refer to all the participants. Doing so could help avoid the single male participant in the project being

identified since there were few males in the institution. The code sheet matching pseudonyms to real names is kept in my laptop with password protection. I informed the participants that they would not be identified in any part of the study to protect their privacy and right to anonymity. Confidentiality encourages participants to speak openly and in good conscience (Simons, 2009). Moreover, to ensure confidentiality, transcripts of moderation discussion, verbal reports and interviews are secured in password protected files and folders in my password protected laptop and portable storage drive to ensure that only I can access the data.

In addition, my previous professional roles, involving training raters and monitoring rating quality may have been a concern to the participants as they may have been afraid that their participation in this research project might have been judged, thus affecting their job. This could have led to them not being willing to discuss openly the related issues. Being aware of this potential concern, I explained to the participants that my study did not aim to judge the quality of their rating job under any circumstances. The aims of the study were to further an understanding of the rating process and their rating practice development. Therefore, their personal opinions and their experiences of being VSTEP raters were highly valuable to the study. I ensured that their information would not be accessed by their current team leaders, the director of the centre and other participants in the context. I also assured them that the generated data would not be handed to or used by any official or unofficial authority which could negatively affect their status or the practices of the participants at any level or under any conditions. Therefore, the participants appeared to be open in discussing their opinions related to the scores they gave in the moderation discussion (further information in section 3.7.1) and in the interview (section 3.7.3). They were also confident in conducting think-aloud protocols after being trained to do so (section 3.7.2). I also ensured that the participants' attitudes,

perceptions and practices would not be mentioned to the other participants or discussed with anyone else.

3.7 Methods of data generation

According to the nature and complexity of scoring decisions, different data-generation instruments were used in the research project. First the participants attended moderation discussions, then they rated 15 speaking performances by using think-aloud protocols (TAPs) and finally they participated in semi-structured one-to-one interviews. These instruments were used collaboratively to answer each research question and gain a comprehensive view of raters' scoring decisions (see Table 3.3). Further explanation and justification of the selection and procedure of employing the instruments are presented and discussed in detail in the following sections.

Table 3. 3: Summary of data generation methods

	Research question	**Data generation method(s)**
1	What are the mental processes of rating speaking performances? - What differences between experienced and novice raters, if any, are seen in attention paid to language features described in the rating scale? - What differences between experienced and novice raters, if any, are seen in decision-making strategies?	Observation of the moderation discussion Think-aloud protocols Semi-structured interview

	- What differences between experienced and novice raters, if any, are seen in attention paid to test-taker proficiency levels?	
2	What are the factors that cause disagreements among all the raters in making score decisions? In what ways and to what extent do these factors affect raters' decision in their ratings?	Observation of the moderation discussion Semi-structured interview
3	In what ways do raters develop their rating practice?	Semi-structured interview

3.7.1 Observation of moderation discussion

This section starts with an explanation of what a moderation discussion is, how significant it is in this study and how the data was generated in this process.

The moderation discussion is an important step in a rating procedure. Its purpose is to monitor raters' behaviours by reminding trained raters of the rating scale before they begin their rating work and illustrating the different levels of the scale in concrete terms. This process is a compulsory component in maintaining score quality by enhancing raters' agreement in scoring (Ga, 2017). The ten benchmark responses are authentic responses that are chosen by the centre to represent certain levels on the scoring guide. The raters are required to listen to those performances and decide scores analytically and holistically for

those examples. After that, they discuss their scores and give the reasons why they have arrived at those scores. The scoring leader, a member of the centre, facilitates the discussion to ensure that every rater is accurately following the scoring guides. The moderation discussion taken in this research project was slightly different from this process: detailed discussion of the differences is presented shortly below.

The moderation discussion was of significance in achieving the aims of the research project as the insights from the moderation discussion helped partially uncover the factors that may affect the raters' scoring decisions through their discussion and negotiation of the meanings of the scores they awarded. As the raters shared what they thought, why they made the decisions of scores and whether they agreed with the other raters' decisions in the moderation discussions, these insights helped find answers for research questions 1 and 2 (see Table 3.3). Being an observer in the moderation discussion could then help me further my understanding of the factors that may affect the raters' scoring decisions by situating people's behaviour within their own socio-cultural context (Hennink, Hutter, & Bailey, 2011). Moreover, Robson and McCartan (2016, p. 320) argue that the way people say may be different from the way they do; thus, observation enables the researcher to look directly at interactions in the context rather than relying on second-hand accounts. This gives observation a high level of authenticity that is rare in studies about raters' behaviours. Among studies investigating raters' behaviours in the literature, Davison's (2004) study was atypical, generating data from the moderation discussion where raters discussed their opinion and decisions in groups.

I decided to be an overt complete observer (Robson & McCartan, 2016), whose identity and the research procedure were explained to the participants. This helped me study people's experiences and reactions in a natural setting

(Emerson, Fretz, & Shaw, 2001) by observing their engagement in the conversation of scoring speaking performances and the way they negotiated their ideas with other raters.

The moderation discussion in my study was conducted with three benchmarking responses, instead of ten (as stated in Ga, 2017) because of time limits. My initial plan was to observe the moderation discussion at two different times because I was aware that there might be time conflicts among the participants and it seemed to be impossible to gather all the participants at one time for the moderation discussion. However, although I was aware of the time constraint, I was not able to conduct the moderation as planned. The first moderation was conducted with 8 participants and was facilitated by a member of the centre as scheduled. For the other 6 participants, I myself had to conduct the session individually at 6 different times as described in Table 3.3 below. All of the moderation discussions were audio-recorded with the permission of the participants.

Table 3. 4: Summary of moderation schedule

Moderation	**Participants**	**Facilitator**
1	8 raters	A member of the centre
2	1 rater	The researcher
3	1 rater	The researcher
4	1 rater	The researcher

5	1 rater	The researcher
6	1 rater	The researcher
7	1 rater	The researcher

During the first moderation, being an overt complete observer, even though the event was recorded, I took notes of the participants' behaviours, actions and interaction within the context. The participants listened carefully to the instruction of the facilitator and to the three recorded performances. After each performance, the facilitator asked for the scores the participants arrived at for each rating criterion and the reasons why they decided the scores. The participants were calm in explaining their decisions of scores even though they were aware that their scores were sometimes different from each other.

As I was aware that I would have to conduct the moderation for the other raters, I took careful notes of what was said by whom and how it was said in the first moderation so that I could give an accurate description of what actually happened to those who did not manage to be present in the first event. In the individual moderation, I took up the role of facilitator of the moderation discussions by playing the benchmark speaking performances to the raters and asking for their analytical and overall scores for each performance. I listened to their explanation of how they arrived at the scores. After that, I described in detail what the other raters had thought in the first moderation and asked if they agreed with the comments and why they agreed or disagreed, and finally I presented the final scores decided in the first moderation and asked what they thought about it.

The data generated from the individual moderation might have been, to some extent, influenced by what I presented to them in several ways. The individual rater might not have been confident to provide their opinion when knowing that there were 8 raters participating in the first moderation. Moreover, as an observer, I might have brought my own perceptions and/or interpretation of the event while providing the description of the first moderation, thereby possibly affecting the individual rater's perception of the information. In order to minimise these potential influences, I tried to describe in detail what each rater in the first moderation said by giving a direct quotation in a neutral voice tone and eliciting the individual rater's thinking by asking what she thought of the comments and if she agreed or not and why. Moreover, I tried to reassure the participants by saying that their opinions were highly important and were allowed to be different from those of other raters.

However, there were differences regarding the engagement level of the participants in the group moderation discussion and the individual ones. It was observed that the 8 participants in the group moderation discussion seemed to have more opportunities to interact with each other, resulting in richer data, compared with that from the individual moderation discussions. For example, the ideas raised by one rater were either supported or rejected by the other raters, which allowed more similarities and differences to surface in the group discussion. In contrast, in the individual discussions, the individual rater provided their responses to the ideas raised in the group discussion without real interaction with the other raters. This could, to some extent, have lessened the richness of the data provided even though the raters in the individual discussions were able to explain the reasons behind their decisions of scores and their thoughts about the ideas raised in the group discussion.

3.7.2 Think-aloud protocols

Think-aloud protocols (TAPs) is a long-established method in psychological and social research and that has been widely used in rater cognition literature (used in Cumming et al., 2002; A. Brown, 2006, etc.). This method requires participants to speak out loud their thoughts while performing a specific task. There are two main variations in the way TAPs are conducted, namely concurrently and retrospectively (Suto, 2012), but according to Kuusela and Paul (2000), concurrent think aloud protocols impose less cognitive strain on participants' memory, thus they are more effective in providing richer information about the thinking process than retrospective think aloud protocols. Gilhooly and Green (1996) also favoured this type of 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). Thus, regarding my research project, which aimed at furthering understanding of the rating process of VSTEP raters, particularly the features of scoring behaviours (see Research question 1 in Table 3.3), TAPs seemed to be the most suitable way of generating data in this respect.

One of the most common criticisms of this method is that it tends to provide unnatural evidence of 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, I conducted the TAPs in two different times for each participant with a time gap of between 1-2 weeks so that the participants did not forget the TAPs procedure and they would become familiar with the TAPs, thus being closer to the nature of the process. One participant (Lilac) was not able to conduct the TAPs due to time conflict.

Planning and design

Before TAPs were conducted, several issues suggested by Green and Gilhooly (1996) were considered carefully, including available resources, feasibility, practicality, the number of protocols needed, a plan for the data analysis, required equipment, cooperation and the motivation of the participants.

I was aware that conducting and analysing protocols is time-consuming, so I tried to generate data from protocols as soon as I could. I contacted the participants in Vietnam to book appointments with them early in order to avoid potential delays. I also contacted the centre to book the use of the cassettes, recorders, and a quiet room during the time of the data generation process.

Regarding feasibility, all of the participants had experience of teaching English as a foreign language, so I assumed to some extent that they were familiar with thinking out loud as at some point in teaching they need to demonstrate to their students what reading a text means to them or the listening procedure and how to get a right answer. I also provided training to familiarise the participants with TAPs.

Practicality involves the question of whether it is necessary for the researcher to be physically present. In this study, the participants recorded the material themselves since self-recording by participants was sufficient for what was being studied as prompting and observations of bodily movements were not

necessary. Moreover, the instruction sheet (Appendix 5) was given to the participants to ensure that they remembered the steps without my presence.

The question of how to motivate the participants was taken into great consideration. Having understood that when they are tired or in a hurry, they may provide incomplete accounts of their thoughts, I contacted the participants and booked potential dates for the two TAPs sessions. I also informed them of the possible duration each session may take so that they were aware of the time and were able to arrange their work. For some participants who proposed some dates and I thought it might affect their upcoming job, for example their classes started at 1pm and they proposed they may come at 10am. In such cases, I asked for another possible date as I did not want them to be in a hurry in doing the TAPs as it may take longer than expected. Moreover, I prepared several bottles of water and snacks in the room so that the participants could have them when needed.

Preparation of subjects

Practice of the tasks that the participants are required to do is highly important, as suggested by Richardson (1996, pp. 53-58) since the training allows the participants to be familiar with the required procedure and to clarify any misunderstandings. I adopted the procedure of training suggested by Ericsson and Simon (1993, pp. 16-18) with 4 tasks described in Table 3.5 below.

Table 3. 5: TAPs training procedure (adapted from Simon, 1993)

Brief introduction to Think-aloud protocols (definition, purpose, and the process)	Raters can use any language that they feel comfortable with (either English or Vietnamese)

Training task 1	Describe your favourite room in your house
Training task 2	Describe the way from your home to this place today (including as much detail as possible)
Training task 3	Watch a 2-minute video, and tell me what you are thinking about. Please feel free to stop the video at any time you want.
Training task 4	Listen to the first 2 minutes of a VSTEP speaking performance and practice TAPs

The first two tasks are relatively simple to familiarise the participants with the first experience of thinking aloud and relax them if they are nervous about the task (Richardson, 1996). The third task is to check if the participants actually think aloud. One of the participants seemed shy while performing this task. It was almost impossible for her to speak aloud her thoughts. I tried to explain the procedure again to her, encouraged her to speak aloud and selected two more videos for her to practice until she felt more comfortable performing the task. The last task resembles the real task that the participants are expected to perform. Any misunderstandings were clarified after doing this task. For several

participants, more minutes of the performance were given to practise until they felt confident to conduct the tasks. The training was conducted individually and before the first TAPs were conducted.

Recording of verbal reports

The participants were seated in a quiet room in an authorised area of the centre in order to avoid unnecessary interruption by "other voices or external disturbances" (Richardson, 1996, p. 59). They were given a cassette (with a pair of earphones when needed) to listen to the performances, a recorder to record their verbal report, and an introduction sheet to remind them of the aims of the task and the necessary steps of performing the task. The recorders were fully charged and checked by the researcher before being given to the participant. Each participant was given one recorder for the whole period of the data generation process. The recorders were kept in a locker after each use to ensure the confidentiality of the data and the participants. Table 3.6 below summarises the equipment the raters needed.

Table 3. 6: Summary of necessary equipment for the raters

Equipment	Notes
CD players with a pair of earphones	To listen to the recorded speaking performances
CDs	Recorded speaking performances
Recorders	To record the participants' verbal report

Information sheet	To remind them of the expectation and procedure of the task

3.7.3 Interviews

This study relied on interviews as one of the main data-generating techniques to gain rich information from the participants about their experience of scoring speaking performances in a high-stakes test in Vietnam.

Interviews enable participants to discuss their interpretations of the world in which they live in, and to express how they regard situations from their point of view. Thus, the interview is not simply concerned with collecting data about life; as Cohen et al. (2011) argued, it is part of life itself. In other words, the meaning-making process of the participants is the central part, which can be accessed via interviews; thus, interviews enable me to get closer to the personal world of the participants. This epistemological point of view is in line with my research's theoretical perspectives since interviews can be considered as the heart of human interaction for producing knowledge and emphasising the social context of the research data (Kvale, 2007).

Among the three types of interview (structured, semi-structured and unstructured), I adopted the semi-structured one because of its flexibility and interactiveness. In the semi-structured interview, the researcher has a specific topic to learn about, prepares a limited number of questions in advance and plans to ask follow-up questions (Rubin & Rubin, 2012). The flexibility in semi-structured interview encourages the interviewee to answer at length and in vivid detail (Rubin & Rubin, 2012) and it makes them feel as though they are leading and dictating the pace of the conversation. Consequently, the researcher could obtain an in-depth understanding of the research phenomenon. Besides, this interview strategy is more of a guided conversation and such kinds of interaction

are important when interviewing raters about their scoring experience in which there might be a number of factors coming into play, namely opinions, reasons, feelings, beliefs, etc. This aided the researcher to partly find answers for all of the research questions (Table 3.3).

All of the interviews in this study were conducted in the mode of face-to-face and one-to-one semi-structured interviews because this format allowed me to discuss the issues in detail with the participants. Moreover, it enabled me to notice the participants' unobservable aspects such as feelings, thoughts and intentions, and behaviours, and reflect how they make meanings of the world and their practices within it (Henn, Weinstein, & Foard, 2009), which added significant meaning to the data. One-to-one interviews were important as the interviewees felt free and comfortable and were therefore willing to discuss any personal factors relevant to the study. They had freedom to discuss their perceptions and attitudes to the scoring experience they had encountered.

The semi-structured interview questions (see Appendix 6) were designed to explore the detailed account of scoring speaking performances in a high-stakes test in Vietnam and the way that rating practice develops. The set of interview questions was developed to reflect the prior themes identified from reviewing the existing literature, with key themes of the research aims and research questions. The focal content of the interview concerned:

- Factors that affect their scoring decisions (rating scales, teaching experience, scoring experience, professional knowledge, perceptions of a good speaking performance, political conditions)
- The way their rating practice has been developed (similarities and differences between the way they rated before and now; the strategies/process that help them to be more consistent in marking)

The interviews were conducted after the moderation discussion had been observed and the TAPs were completed since it gave me the opportunity to further the understanding of their beliefs and practice in scoring as well as giving the interviewee the chance to reflect and discuss their ideas and opinions about their beliefs and practices. With the participants' permission, all the interviews were audio recorded, which enabled me to obtain a full transcript of the interview. However, during the interview I did take notes to help me formulate new questions and facilitate analysis, for example in locating important quotations from the tape (Patton, 2002, p. 383). One participant (Lavender) was not able to attend the interview due to her emergent health conditions.

Even though the English proficiency level of all the participants is C1 level and above (according to CEFR), as this is one of the requirements to be a VSTEP rater, all the data collection methods were conducted in Vietnamese with some code switching when some key terms were mentioned. I decided to use their own L1 language because of a number of reasons. First it could help gain access and create trust and rapport with the participants, thus encouraging participants to talk freely in ways that can reveal the distinctiveness and complexity of their perspectives that the researcher seeks (Hammersley, 2008). Moreover, fluency in their own language allowed the researcher to notice nuances of expression, tone and body language which the use of a second language might inhibit. Third, familiarity with the communicative norms of society (Briggs, 1986) might help the researcher be aware of the underpinning culture in order to study most naturally the individual lived experience (Brinkmann, 2013).

3.7.4 The piloting of interview questions and TAPs instruction

Piloting of research instruments is used to examine aspects of a research design, including refining research questions, stimulus data-generation methods, and estimating the time and costs (Robson & McCartan, 2016). This allows the

researcher to check the intelligibility and unambiguousness of the instruments and to make necessary adjustments before the final commitment to the design. Moreover, the piloting stage was beneficial for me since I could familiarise myself with the process involved in conducting an interview and the TAPs training.

I piloted the interview questions with two VSTEP raters in Vietnam to check the clarity and the order of the questions. The interviews were conducted through Skype, an internet-based software allowing good quality voice transferring. The piloting confirmed that the questions were clear and answerable. From the piloting, I perceived that it is important to allow silence in the interview to gain additional data from the interviewee and that 60 minutes was an optimal time for the duration of one interview. Regarding the TAPs training, I piloted it with a member of the centre, which confirmed that the instructions were clear and the tasks were helpful in familiarising people with the process required.

3.7.5 Timeline of the research project

This section provides the timeline of the stages of completing the study which was conducted (Table 3.7).

Table 3. 7: Timeline of the research project

Time	Tasks
18/09/2017 – 17/12/2017	• Select and justify the research topic • Do extensive reading, identify the research gaps, and generate research questions • Select and justify overall research design • Complete the research proposal

18/12/2017 – 17/03/2018 18/03/2018 – 17/06/2018 18/06/2018 – 17/09/2018 18/09/2018 – 17/12/2018	• Review the literature of rating research • Complete the progression report year 1 • Generate data in Vietnam • Transcribe the generated data
18/12/2018 – 17/03/2019 18/03/2019 – 17/06/2019 18/06/2019 – 17/02/2020	• Analyse and interpret the generated data • Complete the progression report year 2 • Analyse and write the findings
18/02/2020 - 17/07/2020 18/07/2020 – 17/10/2020 18/10/2020 – 17/03/2021	• Update and revise literature review chapter • Complete thesis introduction and conclusion • Complete the first draft of the whole thesis
18/04/2021 – 17/09/2021	• Revise and Submit the thesis

3.8 Data analysis

Interpretative phenomenological analysis (IPA)

The aim of IPA is to provide detailed examinations of how participants make meaning of their personal and social world, including their perceptions and responses to particular experiences, and/or events (Smith & Osborn, 2008). IPA is well suited to my study since my study aims to understand in detail the scoring decisions of VSTEP raters and the development of their rating practice from their own perspectives. It attempts to "understand what personal and social experiences mean to those people who experience them" (R. Shaw, 2010, p. 178). In other words, I was trying to access the participants' inner world through the participants' lens at a particular event (VSTEP speaking test) in a particular place (Vietnam) and time. Thus, in Smith and Osborn's (2008) terms, the analysis

is phenomenological as it "is concerned with trying to understand what it is like, from the point of view of the participants, to take their side" (p. 53). The in-depth analysis of the data following IPA would help me unpack the complicated chain of connection between people's talk and their thinking and even emotional state in order to understand what raters are thinking and feeling while making scoring decisions and how their scoring practice has developed through their own lens. Studies of these have rarely been found in the literature since mainstream language testing is still strongly committed to "highly technical" methodology (McNamara, 2009, p. 608), unpacking these issues through numbers rather than the participants' own perspectives.

At the same time Smith and Osborn (2008) also emphasised the active role of the researcher in their interaction with their participants and with their data during the research process. Thus, IPA also involves making analytic interpretations about the researched experiences and about the person who experiences it. In other words, a dual interpretative process is at work, which is known as the *double hermeneutic*: The participants are trying to make sense of their world and the researcher is trying to make sense of the participants trying to make sense of their world. Hence, a detailed IPA analysis can also involve asking critical questions because access to the participants' inner world "depends on, and is complicated by, the researcher's own conceptions; indeed, these are required in order to make sense of that other personal world through a process of interpretative activity" (Smith & Osborn, 2008, p. 53). This is concurrent with my epistemological stance which I have explained in section 3.2. Although the meaning-making process of the participants is central to phenomenological approach, to get close to the participants' personal world, a researcher needs to engage in an interpretative activity. In this sense, understanding is understood as identifying or empathising with and trying to make sense of. I agree with Smith and Osborn (2008) who argued that "allowing

for both aspects is likely to lead to a richer analysis and to do greater justice to the totality of the person" (p. 54).

In this study I followed the six steps of IPA developed by Smith et al. (2009) as explained in Table 3.8 below. IPA was employed to analyse the generated data, including data from the observation of the moderation discussion, the TAPs and the interviews. Although I was aware that IPA is not a common method used in analysing data from TAPs, there are multiple ways to read a text, as suggested by Lumley (2005) and Paltridge (1994). IPA was helpful in TAPs analysis for its clear guideline and its embrace of details. The following section demonstrates what the researcher did for all the steps.

Table 3. 8: IPA steps (Smith et al., 2009)

Step	**Description of the process**
1 – Reading and re-reading	Reading and re-reading the transcript, record some of the own initial, and most striking observations about the transcript
2 – Initial noting	Note anything of interest within the transcript. The exploratory comments can be descriptive comments, linguistic comments and conceptual comments
3 – Developing themes	Analyse exploratory comments to identify themes

4 – Searching for connections across themes	Develop a charting, or mapping of the themes to point to all of the most interesting and important aspects of the participant's account
5 – Moving to the next case	Move to the next transcript and repeat the process
6 – Looking for patterns across cases	Show how themes are nested within super-ordinate themes and illustrate the theme for each participant

Step 1 – Reading and re-reading

The first step of an IPA analysis involves immersing oneself in some of the original data. I started by listening carefully to the recording and reading the transcription both in Vietnamese and English several times. All the data was generated in Vietnamese for the reasons discussed in section 3.7 and all the transcription was translated into English for reasons related to data quality (explained in section 3.9 – Quality of the data). While listening and reading the transcript for the first time, I noted down my thoughts, feelings and what reminded me of the emotions and the expressions of the interviewees in my research journal (Brinkmann, 2013) (a sample of this is provided in Appendix 8).

Step 2 – Initial noting

This step examines semantic and language use on a very exploratory level (Smith et al., 2009). The researcher maintains an open mind and notes anything of interest within the transcript. This process ensures a growing familiarity with the transcript; moreover, it begins to identify specific ways in which the participant

talks about, understands and thinks about an issue. I started writing notes on the transcript as I started reading, and further exploratory notes or comments were added with subsequent readings.

I tried to identify and describe the things which matter to the participants such as key objects of concern, for example relationships, processes, places, events, values and principles, and the meaning of those things for the participant, including what those relationships, processes, places are like for the participant. I tried to understand the participants' inner world. I looked at the language they use, thinking about the context of their concerns (their lived world), and identifying more abstract concepts which can help to make sense of the patterns of meaning in their account. In the exploratory notes, I tried to include three types of comments suggested by Smith et al. (2009):

- Descriptive comments: focused on describing the content of what the participant has said, the subject of the talk within the transcript
- Linguistic comments: focused on exploring the specific use of language
- Conceptual comments: focused on engaging at a more interrogative and conceptual level

Table 3.9 – column 2 (exploratory comments) illustrates my initial notes. Some interpretation developed at this stage unavoidably drew on my own experiential, professional knowledge and/or my own pre-understandings in order to sound out the meanings of key events and processes for the participants (Smith et al., 2009). For example, the comment "acceptance of being inconsistent" and the interpretation of "that is the mistake not to make" was drawn on my knowledge of common rater errors. However, I always tried to ensure that "the interpretation was inspired by, and arose from, attending to

the participant's words, rather than being imported from outside" (Smith et al., 2009, p. 90).

Table 3. 9: Sample of Initial noting and developing emergent themes

Transcript of interview data of Rose	Exploratory comments	Themes
24:45 **Can you estimate when you can give a final score for a TT?** I give the final score in the end when saying goodbye, but I often think back. I know I am wrong when comparing TTs sometimes, but it's inevitable. And in many occasions, my sense might not be right, like it may not be exactly 5 or 6, but something in between 5 and 6, or sometimes 6.5. It's different from marking writing in the sense that when marking writing, I have time to consider, and an easy-to-read writing can be marked faster than the one that is unstructured, or short, or off-topic. But sometimes I still need to reread it in case I did not fully understand their ideas. With speaking test, I mark the same, but when the TT's performing time is over, I look back at my notes so that I can carefully weigh all the criteria, either up or down, of course it will not exceed that band, but the score for some criteria could go up to some point.	Re-thinking after giving the final score Acceptance of breaking the principle (the mistake not to make) and consider it as inevitable. Human factor. Mentioning of own sense while being confused of giving 5 or 6 or between Compare time for scoring between writing and speaking. Scoring time in speaking is much less Notes seem to play an important role in deciding the final score Careful consideration of score for each criteria even within the level	Tension between knowing and doing The own sense/scale descriptors fail to cover the complexity of performances Time constraints Note-taking
26:11 **When was the last time you marked VSTEP test?** Last Sunday. I went to [Name of the place]. **Can you describe your marking procedure starting from when TT entered the room?** When a TT just enters the room, the recorder is already ready and will be pointed at the TT. After greeting, in order not to waste time, I don't allow them to write their name immediately, I just ask them to sign it. I would write it later, because 10 minutes is not much, but many people might want to express themselves, so I don't want to waste their time. I always bring two things, smartphone, ah no I forgot we are never allowed to bring mobile phones, my mistake – it's the recorder. Many people never estimate time, but I always do, even one minute should be exactly 1 minute. I noticed many people just sit still for 2 minutes, and 2 minutes	Awareness of the significance of time to the TT	Time awareness

pass very quickly. In the first part, I tried to ask both sections, and not focus too much on section 1 or section 2. It is very typical of me to never change questions' order, while many people like to ask topic 2 first. It is similar to the past when I took the exam, I tried to do from beginning to end so that the reader has the feeling that I could do them all. Sometimes I couldn't be able to do all, but always in order. I also watched the time while I was listening and taking notes. For example, if I let part 1 last for 4 minutes, it will take the time of part 2. Part 1 is quite simple, which TT at band B1 should be able to answer, while I feel the level of difficulty increases in part 2. And in part 3, the way TTs develop and connect ideas, and criticize are evaluated from B1 to C1, instead of a specific band. I always watch time during part 1. It's impossible to be exactly 3 minutes, because in many cases I don't want to interrupt while they are speaking, but I never let it exceed 4 minutes. After that I move to part 2, I will tell them to move to part 2, then give them scratch paper, and they already have pen. I don't watch the recorder because it shows the time, for some recorders the time only comes out when you press the button, so I look at the watch, which has four hour markers 12, 3, 6, 9. I can't remember the position of 25 because it's very difficult to count to 25, so I estimate in between 3 and 6, then when it returns to the same number, 1 minute is up. After 1 minute, I ask them to speak, and after they give all reasons, I'll ask questions related to those choices, then a critical question. Sometimes they misunderstand, I'm not going to interfere with their understandings whether it's right or wrong. Then following my script, I will say I agree with their choice. Then moving to part 3, the introduction is too long, in the beginning, I tried to read, but it depends on time. If I don't have time, I'll read connecting the sentences together. Instead of using two simple sentences, I connect using participle phrase or compound sentence to make it shorter. Many people don't really pay attention, they just wait for me to finish talking to take notes. I continue like that for one minute, then let them start speaking. Some can speak well, some can't speak, and some are very good at elaborating questions and giving ideas. Then I watch the clock again, and take notes as well. I take notes and watch the clock at the same time to see the remaining time for round-off questions, which are usually 2 or 3 questions. Time is normally up after one or 2 questions are asked.	Own strategies of managing time during the test to maximise talking time of the TT and strictly follow the instruction in the exam script too Strictly follow the exam script. Why does she follow the script strictly? Is it to ensure fairness for TTs? Take notes	Time management Exam script bound/Fairness Note taking

Step 3 - Developing emergent themes

The main task in turning notes into themes involves an attempt to produce a concise and succinct statement of what was important in various comments attached to a piece of transcript. Themes are usually expressed as phrases which speak to the psychological essence of the piece and contain enough particularity to be grounded and enough abstraction to be conceptual. The focus is on capturing what is crucial at a point in the text but inevitably it will be influenced by the whole text as Smith et al. (2009) argued it is "the hermeneutic circle where the part is interpreted in relation to the whole; the whole is interpreted in relation to the part" (p. 92). See Table 3.9, column 3 which presents the emergent themes for the extract of the transcript in column 1.

Moreover, the themes reflected not only the participant's original words and thoughts but also the researcher's interpretation. For example, the first emergent theme, *tension between knowing and doing*, captures the initial exploratory notes relating to Rose's acceptance of making mistakes (comparing TTs with TTs) and her argument that it is inevitable. The theme, therefore, related directly to the content of Rose's talk; moreover, within the theme title, *the influential factors* reflected the researcher's interest in the psychological construct of the participant. Again, this process is congruent with the theoretical stance I adopted in this research.

Step 4 - Searching for connections across emergent themes

In this step, I reviewed all the themes in chronological order and moved the themes around to form clusters of related themes. This allowed me to point to all of the most interesting and important aspects of the participant's account. Moreover, I used tables to sort the themes which are relevant both to the themes identified in the literature and my research questions. When I had a

collection of TT themes and sub-themes, and all extracts of data that have been coded in relation to them, I started analysis with the theme from my research questions. For example, Table 3.10 illustrates how the themes from the data extract in steps 2 and 3 are sorted and classified under the names of *Stages of rating practice development* and *Influential factors,* which are two key issues in my research questions. I kept reviewing the themes, clustering themes and connecting them with my pre-understandings of the literature and my research questions in order to show the interconnections between recurrent group themes.

Table 3. 10: The development of a super-ordinate theme

Stages of rating practice development	Influential factors
Time constraint Time awareness Time management Note taking Exam script bound	Tension between knowing and doing Their own sense?/ Trusting the self Problems with scale descriptors

It is important to note that there were three sets of data for each participant, including the recorded moderation discussion, TAPs and interview. Each set of data was analysed separately in the first three steps. Step 4 involved identifying connections across emergent themes; therefore, the themes identified in each data set were pulled together under the light of the research questions. For example, RQ1 investigated the mental process that VSTEP raters experienced.

The themes which were related to the stages of rating in the mental rating process were placed together. These themes are illustrated in Table 3.11.

Table 3. 11: Emerging themes for RQ1

Aspects of speaking performances attended	- Grammar - Vocabulary - Pronunciation - Fluency - Discourse management - Others
Allocation of initial scores	- Rating criteria - Time of allocating
Finalising scores	- Matching strategy - Simplifying key terms in the scale descriptors - Referencing to holistic rating - Compensating - Using own sense

NVivo 12 was used from this stage onward to help with themes management (see Appendix 7 for a template of coding).

Step 5 – Moving to the next case

This step involved moving to the next participant's account, and repeating the process (i.e. steps 1,2,3 and 4). Smith et al. (2009) notes that it is important to treat the next case on its own terms, to do justice to its own individuality. This

means bracketing the ideas emerging from the analysis of the first case while working on the second. However, the authors also acknowledged that the process of analysing the next case will be inevitably influenced by the way the first case is analysed. Thus, it was important to allow new themes to emerge with each case and by following systematically the first four steps outlined, there was scope for this to happen. The process of analysis was sequentially applied to the fourteen participant raters.

Step 6 – Looking for patterns across cases

The next stage involved looking for patterns across cases. I laid each table out on a large surface and looked across them. I tried to answer the following questions:

- What connections were there across cases?
- How does a theme in one case help illuminate a different case?
- Which themes are the most potent?

Since my sample included 14 participants, which was a large corpus, measuring recurrence across cases is important (Smith et al., 2009, p. 106). Table 3.12 below illustrates how the recurrent themes of the participants' attitudes towards the VSTEP test were counted.

Table 3. 12: Recurrent themes

Participants	"A jigsaw test"	"Appreciation of a home-grown test"
Daffodil	Yes	Yes
Orchid	No	Yes
Tulip	No	Yes
Sunflower	Yes	Yes
Lotus	Yes	No
Daisy	Yes	Yes
Lavender	NA	NA
Rose	Yes	Yes

Hyacinth	Yes	Yes
Jasmine	Yes	No
Lily	Yes	Yes
Iris	Yes	Yes
Peony	Yes	Yes
Lilac	No	Yes
Present in over half of the sample?	Yes	Yes

Table 3.12 indicates the super-ordinate themes that were present for each individual participant and these were then calculated to illustrate whether the themes were prevalent in over half of the cases. This decision of counting over half of the cases was taken in this research project because it allowed for a balance of the relationship between convergence and divergence, commonality and individuality. This way of counting recurrent themes allowed the researcher to "retain an idiographic focus on the individual voice at the same time as making claims for the larger group" (Smith et al., 2009, p. 107).

After counting the recurrent themes, a master table of themes for the group was formed (Table 3.13). These main themes (A, B, and C) were elaborated into three different chapters (Chapters 4, 5, and 6). In order to illustrate the lived experience of the participant raters vividly, each chapter starts with a vignette which is generated from the data to narrate different aspects of the experience the raters encountered. The aspects illustrated in the vignettes were the data "taken to be representative, typical or emblematic" (Miles & Huberman, 1994, p. 81) and are presented as a "bit of a story" (Thomson, 2017). The vignettes are important to the analysis since they involved in a systematic process of moving from themes which emerged across multiple individuals to a single representation of these.

Table 3. 13: Master table of themes

Master table of themes for the group

A. The scoring process as experienced by the participant raters

   Speech features attended

   Initial score allocation

   Decision-making strategies

B. Factors causing disagreement among all the participant raters

   Global and local dimensions

   Orientation towards rating criteria

C. The way the raters developed their rating practice

   The context:

   - A jigsaw test – the trust between localised test and international standardised tests
   - Appreciation of a home-grown test
   - The significance of being VSTEP raters

   Stages of development:

   - Feeling overwhelmed at managing multi-tasks at the same time
   - In more control of doing the tasks required
   - Knowing what to do with confidence

3.9 The quality of the data

The qualities of validity and reliability were originally established in natural sciences (Seale, 1999). However, their relationship with natural sciences and the different epistemological basis of qualitative inquiry led to applying these values to determining the quality or viability of qualitative evidence (Ritchie & Lewis, 2003). In qualitative inquiry, validity is defined as how accurately the account represents participants' realities of the social phenomena and to what extent it is credible to them (Schwandt, 1997). In other words, validity can be assumed to refer not to the data but to the inferences drawn from them (Hammersley & Atkinson, 2007). Various authors have constructed diverse typologies of validity: authenticity, goodness, verisimilitude, adequacy, trustworthiness, plausibility, validity, validation, and credibility (Lather, 1993; Lincoln & Guba, 1985; Maxwell, 1996). In this study, I follow the criteria constructed by Lincoln and Guba (1985), which are listed below, to ensure and enhance the quality of my research project:

- Credibility
- Transferability
- Dependability
- Confirmability

The four criteria will be discussed in the following section to explain the different strategies I used throughout the research process.

Credibility in qualitative research deals with the question "How congruent are the findings with reality?" (Merriam, 2009). Different techniques can be used to establish the credibility and trustworthiness of research, namely prolonged engagement, persistent observation, triangulation, referential adequacy, peer debriefing and member checking (Lincoln & Guba, 1985). This study used several

techniques that are relevant such as prolonged engagement, triangulation and thick description.

Prolonged engagement enables the researcher to gain an adequate understanding of an organisation and to establish a relationship of trust between the researcher and the participants as it can enable researchers to further their understanding of the cultural/social setting (Willis, Jost, & Nilakanta, 2007, p. 221). Creswell and Creswell (2018) argued that "the more experience that a researcher has with participants in their settings, the more accurate or valid will be the findings" (p. 201). Although I am currently studying in the UK, I had 9 years' experience of working in the institution; therefore, I possessed an in-depth understanding of the system and "could convey detail about the site and the people that lends credibility to the narrative account" (Creswell & Creswell, 2018, p. 201). This is one of the benefits of being an insider, which I shall explain further in the next section (3.10 - The role of the researcher). There might have been changes during the year in which I was away from the institution; however, a good relationship of trust with the participants was still maintained. Those who agreed to participate showed their enthusiasm and responsibility during the whole period of data generation. Because of their tight teaching schedule on weekdays, some even offered to do the tasks during weekends. Several participants could not come to the scheduled sessions because of family commitments; they were happy to reschedule it. Moreover, they told me they agreed to participate because they were interested in my research project and offered me the opportunity to contact them if I needed further help or explanation regarding the data.

Triangulation is a validity procedure where researchers search for convergence among multiple and different sources of information to form themes or categories in a study (Creswell & Miller, 2000). Concerning this study,

triangulation was achieved by combining three different data collection instruments, namely recordings of moderation discussion, think-aloud protocols (concurrent verbal report), and interviews. These methods provided verifying evidence to locate major and minor themes, which allowed sufficiently deep understanding and insightful interpretations of the lived experience of the participants. These methods provided opportunities to enhance deeper insights in the meanings of participants' lived experience of scoring speaking performances and their rating practice development. Credibility was therefore enhanced by relying on multiple sources of evidence, rather than on the use of a single data source.

One of the most important means for achieving credibility in qualitative research is thick description as Geertz (1973) argued that any single behaviour or interaction, when extracted from its context, could mean a number of things. Therefore, thick description requires that the researcher account for the complex specificity and circumstantiality of their data. In this study, I provided vivid details about the context of the study, the participants, the setting where I generated the data, and a detailed description of each step which helps in transporting the readers into the setting or situation. In this sense, it promotes credibility as it assists the researcher in interpretation and makes the phenomenon clear to readers (Shenton, 2004).

The second criterion is transferability, which refers to a "process in which the reader of the research uses information about the particular instance that has been studied to arrive at a judgement about how far it would apply to other comparable instances" (Denscombe, 2014, p. 299). Phenomenology does not allow generalisation since the findings are specific to a small number of individuals and a particular setting. Moreover, the aim of the qualitative research is "to allow for transferability of the findings rather than wholesale

generalisation of those findings" (Pickard, 2007, p. 20). In phenomenological study, the reader can make links between the analysis, their own personal and professional experience, and the claims in the extant literature (Smith et al., 2009). It is advisable that a rich, transparent and contextualised analysis of the accounts of the participants be provided. This should enable readers to evaluate its transferability to persons in contexts which are more, or less, similar. As I explained above, I provided sufficient and rich description of the context, the sample, the methods which were used to generate the data and the themes which allow the readers to make decisions about the applicability of the findings to other similar contexts or settings (Creswell & Miller, 2000). Moreover, all of my data, which was obtained in Vietnamese (the reasons for this were written in detail in section 3.7 - Methods of data generation), was translated into English by a translation agency with the financial support of the assessment research award I received from the British Council. The translation would help to build an accessible archive, which enables the reader of the research to make decisions about the accuracy of the findings (Creswell, 2014) and the applicability of the findings to other similar contexts or settings (Creswell & Miller, 2000). This helped increase the transferability of the data.

Another criterion to enhance the quality of the data is dependability, which is concerned with researchers' responsibilities for making certain that the findings of the study can be repeated with the same or similar participants (Denscombe, 2014). However, the nature of the phenomena scrutinised by qualitative researchers renders such provisions problematic in their work. In order to avoid this limitation, dependability still can be further enhanced through several techniques. First, it can be achieved through keeping a detailed record of the study journey (see Appendix 8) and of the data generation and analysis process and decisions about the research (Creswell, 2013). Second, I used a triangulation of methods by combining different research instruments to enhance the

trustworthiness of the research (Lincoln & Guba, 1985). Third, I provided a detailed report of the analysis strategies, which allows the reader to gain a clear and accurate picture of methods used in this study.

The last criterion according to Lincoln and Guba (1985) is confirmability. It refers to the degree to which the presented findings reflect the ideas and experiences of the participants, rather than the preferences and characteristics of the researcher (Patton, 2015). Therefore, the role of triangulation in promoting such confirmability must be emphasised to reduce the effect of investigator bias in the study. Moreover, I kept a reflexive diary in which I noted the potential influence that I might bring into the data generation and analysis process. I also kept in mind the role of the researcher including reflexivity (see section 3.10 below) during the study. This self-reflection and reflexivity created "an open and honest narrative that will resonate well with readers" (Creswell & Creswell, 2018, p. 200).

3.10 The role of the researcher

This study adopts insider research which has been defined as the study of one's own social group or society (Loxley & Seery, 2008). In conducting the research project, I took the perspective which "looks at things through the eyes of members of the culture being studied" (Willis et al., 2007, p. 100). In other words, a better way of understanding every dimension of a culture is by dwelling within that culture. In addition to my position as a researcher, I was known by some participants as their former deputy head of one division in the institution where the research took place, as I used to work there before I became involved in the work related to VSTEP tests in the Centre. The others knew me as a test developer and a trainer of the VSTEP rater training program in the study context. Therefore, I had a number of shared experiences with the participants, which defined me as an insider of the researched context. This insider role may have

given me an advantage over an outsider as alongside my knowledge of the Vietnamese language, my understanding of Vietnamese culture and the setting and my prior experience of working with them I could make better choices about the elements to be discussed and examined.

However, I was fully aware that my personal experience and preconceived ideas could also have the potential to affect interpretation of data and introduce bias. Therefore, critical scrutiny was employed to minimise the effects. I chose to be open to equalising the relationship and undermining power as Seidman (2006) argued that equity is essential for building trust and for the participants to be willing to share their experiences. Although I am an insider of the community, I was away for more than one year; thus, it was a good idea to find a way to fit in again. I came to the office where I used to work not only to see my colleagues but also get an understanding of how to present myself properly. Before the moderation discussion and interview, we had a short casual talk about the highs and the lows in our work and life since we last met. I also openly and in a friendly manner answered all of the questions they asked related to my research project. When some of them seemed to worry that their scoring quality would be judged, I tried to reassure them that all the information I got from them was for research purpose only and would be kept confidential; I confirmed to them that their job security would not be in any way influenced. During the moderation discussions and interviews, I tried to be an active listener. I ensured that the participants had enough time to speak openly without being forced to talk about certain points. I showed patience, respect and interest in listening to the participants. I asked them to clarify questions when I thought there was some assumption of the mutual knowledge between myself and the participants. I also avoided answering their questions of related to whether they were doing right or wrong in their responses by diverting them to the importance of their perspectives and behaviours, all of which are valuable. I did

this in order to reassure them and make them more confident and more open in the coming meetings as Glesne (2016) stated that among other qualities that could be used to describe a good researcher is the quality of being reassuring. Finally, I explained how valuable their experience was to me and to my research project and how grateful I was for their participation.

Additionally, I used a reflective diary which "on a daily basis or as needed, records a variety of information about self and method" (Lincoln & Guba, 1985, p.327). I recorded my ideas, thoughts, personal experiences and all the research-related decisions made from the beginning of my PhD journey. For example, before my field trip I was worried that what the participants would say may not correlate with what I have found in the literature and if I should redirect them to those themes. However, Rubin and Rubin (2012) suggested that conversational partners should be encouraged to raise issues that are important to them. The readings encouraged me to think more freely about the people and their issues rather than being bound by my preconceived research structure, which shed considerable light on my research process. The diary also included a chronological record of the events, including things that interrupted the sessions such as a technological problem with the CDs and my feelings and reactions to these (see Appendix 8). Therefore, the reflective diary supported me to "make my experience, opinions, thoughts, and feelings visible and an acknowledged part of the research design, data generation, analysis, and interpretation process" (Ortlipp, 2008, p. 703).

In this section, I have discussed my role as a researcher including awareness of my own predispositions and subjectivity, my role as an active listener, and a learner and how these roles influenced my research process. In other words, I have discussed how I have undertaken my journey through "a self-critical lens"

(Finlay & Gough, 2003, p. ix) by identifying and interrogating personal and professional practices on the part of the researcher.

3.11 Summary

This chapter discusses the philosophical perspectives underpinning the research paradigm and approach used in this study. It also justifies the research strategy, describes the sampling techniques, data-generation instruments, and data analysis techniques applied. In addition, the chapter discusses the issues of ethical considerations, the role of the researcher, and the quality of the data.

This research project is a phenomenological study since it investigates the lived experience of the participant raters when they performed their VSTEP rating job. There were three overarching aims that the project tried to fulfil:

- To further an understanding of the rating processes experienced by Vietnamese teacher-raters in the assessment of speaking performance in English as a foreign language
- To further an understanding of the factors which affect the raters' scoring decisions
- To further an understanding of the ways raters develop their rating practice over time

This study generated data from recorded moderation discussions, TAPs and one-to-one semi-structured interviews. All the 14 participants participated in the moderation discussions in which they explained the reasons for their scores of three bench-marking performances. All participants, except Lilac (due to time conflict) used TAPs to rate 15 VSTEP speaking performances with varying levels of proficiency. The interviews were conducted with 13 participants as Lavender was not able to attend due to her health condition. All the data was recorded with the participants' permission. The participants' first language (Vietnamese)

was used to generate all the data to ensure that the participants expressed their perceptions and feelings freely without pressure and that the researcher could notice nuances of expression, tone and body language which the use of a second language might inhibit.

Interpretative phenomenological analysis was employed to analyse the research data through six steps, including reading and re-reading, initial noting, developing emergent themes, searching for connections across emergent themes, moving to the next case and looking for patterns across cases. Detail of each step was discussed in this chapter. The quality of the data was examined according to four criteria: credibility, transferability, dependability and confirmability. Ethical issues were considered and discussed from the perspectives of the research's ethical approval, access to the research site, informed consent, anonymity and confidentiality, and my professional role. My previous experience of working in the research site gave me several advantages in carrying out this research project, which is also discussed in this chapter. The next chapters (chapters 4, 5, and 6) address the findings in relation to each research question in turn.

Chapter 4: The scoring process as experienced by novice and experienced raters

4.1 Introduction

Vignette 1 – Daffodil (E)

When a test taker comes in, I would, follow the script exactly, welcoming and inviting [the test taker] to sit down. I often introduce my name right away, which also follows the script. After that I ask clearly "What is your full name?" Then I take a look at the ID card and move into part 1. In part 1, I usually ask both of the topics, and rarely ask only one topic, even if their talk is long and I have to interrupt in order to move to the other topic. If the other topic, is too much, I have to ask one or two questions. Perhaps there is not enough time for three questions, but in some cases, they still finish all those 3 questions.

In part 1 I have to remember what they say, whether they can extend their answer, and whether their extension stays on topic, or not. Do they have anything special or interesting? Is there anything right in part 1 that shows a superstar? Generally, when test takers are speaking, I won't look at the rating scale, but I take notes of the ***grammar*** *first. I'll see if test takers use complex structures or simple structures, and how accurate they are, so I normally just use the +/- sign to see in terms of quantity. Then for the* ***vocabulary****, I will see what words they are using. For example, words at level B1, B2 or C1, things like that can all be noted. Or I will see if they make errors about word choices or word forms, whether they have any problems with collocation, less common words or not. For* ***pronunciation****, I also have to assess during the process of taking notes, whether their pronunciation is intelligible or not, whether their individual sounds are accurate or not, stress stops at the level of word stress or sentence stress. Whether they have intonation, have flow? Then when it comes to* ***fluency****, I also search to see if they speak fluently, do they have relative ease with different topics? I check their speaking rate, how their hesitation is, how they extend ideas, or only stop using simple sentences separately. For* ***discourse management****, I'll see how they develop ideas, whether they use examples and give details, or not, the level of appropriateness, whether they use simple or complex connectors, and how they develop ideas.*

After the first part is over, part 1 is done then moving to part 2, I guide them like that, then they have one minute to prepare. Then I'll have one minute to look at the notes to consider which band they are at, roughly. And until then I'll focus on the rating scale. When they start speaking the second part, I still repeat that process for part 2 and part 3. When they finish speaking the third part, I'll make a decision for their final score.

> *But as I already told Thuỷ [the researcher], it's okay for those test takers who have clear characteristics, but there are some who need to be carefully considered. The process of consideration is very painful because when they just stand up and walk out, there will be another test taker sitting in front of me already. And if I don't consider it quickly, the time span in my mind will go away very quickly. Later, even if I look at the notes, I'm not very satisfied with how I eventually gave the final mark. Therefore, I always try to give the mark right after they finish. Because later I'll feel not very confident as to whether I can remember their actual scores or if I gave them an unfair mark.*
>
> *(Interview)*

Daffodil became a VSTEP rater because of a decision made by the faculty in which she was working, in a similar way to other participants in this study. For her, this happened very quickly; the time from the initial decision to attending the first rater training program was very brief. Since then, she had frequently rated VSTEP speaking tests, at least once every two months each year, which means she participated in every test administration. Vignette 1, consisting of several passages extracted from the interview with Daffodil, portrays her rating process at the time of the interview, when she had been doing the job for about 3 years. It is evident in the vignette that her rating process was strongly influenced by her training in general and the rating scale in particular. Vignette 1 introduces this process, which was similarly experienced by other participants, including three main stages: attention given to TT speech features, score allocation and score finalisation.

This chapter, therefore, is concerned with describing the stages of the rating process and the strategies the raters employ to decide scores. It examines in detail the sequence of steps the raters follow while rating, and the attention they pay to the five criteria in the VSTEP rating scale, including Grammar, Vocabulary, Pronunciation, Fluency and Discourse Management. Specifically, this chapter studies in detail:

- the content of raters' comments, and the features they selected for comments during rating (section 4.2);
- the way raters allocated their first scores (section 4.3);
- the way they finalised their scores (section 4.4).

While examining these features, the analysis also compares the rating process of novice raters and experienced raters. Such examination is clearly necessary to unpack the significance of their lived experience of rating VSTEP speaking performances, specifically to answer the first research question "What are the mental processes of rating speaking performances?" and its sub-question "what differences are there between experienced and novice raters in this process?", and to lay out the picture for the next chapter which tries to explore the reasons behind these differences.

Having delineated the structure of this chapter, I present the main findings of the research project which are examined in more depth below. There are two main findings which, I would argue, could be termed significant. First, although all the raters seemed to experience similar stages (attention to speech features, score allocation and score justification) in the rating process, there were differences in rating behavior between novice raters and experienced raters. Particularly, the experienced raters appeared to display more sophisticated attention toward the speech features, which was more evidently seen in higher level proficiency TTs. The explanation for these differences is discussed in detail in chapter 5. This finding contributes to extending the understanding of how experienced raters differ from inexperienced raters, which is an under-researched topic as previously discussed in chapter 2 (sections 2.5 and 2.6). Second, five common decision-making strategies were used by the participant raters. The raters with more rating experience appeared to be more confident in using those strategies than those with less rating experience. I would argue that

the findings presented here offer empirical evidence for the claim that there are no available studies in the speaking assessment literature describing in detail 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). Further findings presented in this chapter include: the significance of the rating scale among the trained-raters, the important role of holistic evaluation and the impact of the scale wordings on the raters.

4.2 Which aspect of speaking performance did the raters attend to?

As Vignette 1 illustrated, the raters generally attended to speech features which were mentioned in the rating scale, including Grammar, Vocabulary, Pronunciation, Fluency and Discourse Management. This section analyses in detail the comments that the raters made while they were rating the speech. This rating aspect was revealed mostly from the TAPs data set, rather than the other data sets (moderation discussion and interview), as TAPs recorded what the raters thought and did while they were rating the speaking performances. Thus, in this section, the majority of the extracts/examples were taken from TAPs; however, a few extracts from the moderation discussion (MD) and interviews were used occasionally where necessary and relevant to help better understand this type of rating behaviour.

The rating process started immediately once the TTs started to talk as the data shows:

Daffodil (E) – TAPs_TT2

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

The TT can use the structure with "because", "and", "when"	*Gr – Complexity*
and there is one grammar mistake related to the use of "very" before main verb.	*Gr – Accuracy*
A bit concerned that the interlocutor asked a question not mentioned in the script.	*Interviewing behavior*
The TT can use the word "balance"	*Vo*
although making a mistake – using "many" with an uncountable noun	*Vo – Accuracy*
The TT can use another complex structure with "because",	*Gr - Complexity*
but wondering if it's due to grammar mistake or pronunciation, I can't identify if the TT uses "to be" in the previous sentence.	*Gr – Accuracy* *Pr - Clarity*
Because I heard the sentence but feel like the TT says "it" then immediately uses an adjective without the appearance of "to be"	*Gr – Accuracy*
The TT seems not…don't know if it's due to ideas or the lack of words…feel like unable to complete the idea.	*DM – Thematic development*
Hesitation… densely occur	*Fl – Hesitation*
The TT uses simple sentences but still makes mistake of using "can" and adjective right away	*Gr - Accuracy*

The example above, from Daffodil, typically illustrates how the rating process started as this process was echoed in the TAPs data of all 13 raters. The first comment made by Daffodil concerned the complexity of grammatical structures that the TT used at the very beginning of her talk. The following comments focused on different aspects of speech rather than one particular aspect. The raters did not pay separate attention to each of the rating criteria when they were rating. This seems to be different from what Lumley (2005) discovered in his study of writing raters. The nature of the speech in speaking performance and the rating time pressure may not allow speaking raters to nominate each criterion at one time whereas writing raters may have more time to reread scripts and refer to the rating scales, which may explain why they could give separate attention to each of the rating criteria.

There are six sub-themes which emerged from the data, five of which are in line with the assessment criteria listed in the rating scale: Grammar, Vocabulary,

Pronunciation, Fluency and Discourse management. Very few comments were identified as “others” when they were not related to the rating scale and these arepresented at the end of this section (section 4.2.6). This section, thus, demonstrates in detail the attention that the raters paid to each criterion, which is supported by the fact that the “others” category where the raters commented on other aspects rather than on the rating criteria was only used in a limited way. Typical examples of how the raters heeded their attention are provided and analysed category by category since this serves as a necessary introduction for Chapter 5, which provides examination of why the raters arrived at different scores for the same TTs.

4.2.1 Grammar

Two major sub-themes emerged under this category which reflected the way the participants made sense of grammar as an assessment criterion: accuracy and complexity of grammatical structures.

All thirteen raters paid attention to how accurately the TT constructed grammatical structures. When the raters identified grammar mistakes, they often repeated what the TTs had just said with emphasis on the mistakes and/or named those mistakes. This behaviour was illustrated by the two examples below, one from Tulip, an experienced (E) rater and the other from Iris, a novice (N) rater.

Tulip (E) – TAPs_TT1

I was improve

It is attract them

Verb tense mistake - “my aunt don’t need”

“Spend for your family” – a basic mistake

“They IS learn”

Children IS – can't understand the complex sentences

I VERY like

Iris (N) – TAPs_TT1

Wrong grammar – "isn't attract"

Wrong pronoun

Subject-verb agreement – grammar mistake

Lack "to be"

The issue of grammar mistakes is one of the central sub-themes shared by all the raters in their rating process. The raters appeared to confidently experience this without much difficulty. Both experienced and novice raters seemed to be able to identify grammar mistakes in speech when they heard them. However, the difference was that those with more rating experience (Tulip, Daffodil, Sunflower, and Orchid) seemed to provide more comments and their comments tended to be more specific. For example, Orchid's typical comments on grammar accuracy were "*this TT has used the simple sentence wrong – 'I like most my'/without a verb*" or "*there are many actions suitable to me – then here a verb is missing*" (TAPs_TT2). In contrast, novice raters tended to provide general evaluative comments showing that the TT was making mistakes, such as "*simple sentences – sometimes have mistakes*" or "*sometimes neglect verbs*" (Peony, N, TAPs_TT2). This characteristic was reflected consistently throughout the data set of grammar.

Complexity of grammatical structures and their range is another important sub-theme which emerged from the data. Generally, all the raters tended to identify not only specific complex structures but also the control of these structures. Before illustrating the examples of this rating characteristic, the discussion of what complex structures of speech were defined as by the raters in the moderation discussions is presented. The differences of definition were revealed

in their score discussion in the second recording in which the raters tried to distinguish band 4 from band 5 of the grammar descriptors. The raters were aware from the rating scale that "the attempt to use complex structures" differentiated band 4 from band 5; however, they seemed not to agree on the types of sentences that belonged to "complex" grammatical construction. The extract below shows the discussion of the issue.

> ***Daisy (N)****: then how do you define complex sentences?*
>
> ***Lavender(N)****: complex [sentences] must have 2 clauses and usually when analysing grammar of the learner they will identify one complex sentence as one main [clause] and one dependant [clause]. If two clauses are connected with 'and', 'but', they are not complex [sentences].*
>
> ***Rose (N)****: I still count compound [sentences] as complex structures rather than simple sentences*
>
> ***Lavender (N)****: it's just a [simple] sentence which is only stretched longer*
>
> ***Hyacinth (N)****: there is only one 'because' sentence in this performance*
>
> ***Lavender (N)****: it is called T-unit, 1 independent clause and all dependent clauses…if the more of them, the broader the T-unit is and the more complex the sentence is.*
>
> *(MD – Recording 2)*

Lavender seemed to perceive the complexity of grammatical structures from the formal view of written grammar while Daffodil and Rose viewed it in a more informal way – the spoken perspective. Lavender was the only rater who used the term T-unit to explain how she perceived complex grammatical structures. This appears to indicate that Lavender relied on a specific aspect of assessing grammar – subordination-based measures (suggested by Wigglesworth, 1997 and Mehnert, 1998) in shaping her perception of assessing speech features and later on her rating behaviour. In contrast, what Daffodil and Rose perceived seemed to be concurrent with Halliday and Matthiessen's (2004) view of speaking in which they stated the ways that speaking is different from written language. Speaking is usually considered less formal in terms of the use of

vocabulary, uses fewer full sentences as opposed to phrases, contains repetitions, repairs and has more conjunctions instead of subordination (Halliday & Matthiessen, 2004). Thus, Daffodil and Rose appeared to use both subordination-based variables and coordination-based variables to evaluate grammatical complexity. These differences in the way they defined "complex structures" were echoed in other individual moderation discussions. The raters who advocated Lavender's view of defining "complex structures" included Lotus, Iris, Jasmine and Peony whereas Orchid, Sunflower, and Daisy were on board with Daffodil and Rose's definition. These differences in viewing syntactic complexity may be indicative of two possibilities. One possibility is that the descriptor of describing grammatical complexity in the rating scale may not have suggested a uniform way of understanding. Alternatively syntactic complexity is a highly complicated construct to be clearly articulated in the rating scale. These differences in perceiving "complex structures" provide empirical evidence to support the claim made in chapter 2 (section 2.8.3) that it is necessary to clarify how raters perceive a complex construct as syntactic complexity in L2 speaking assessment.

Daisy came to her conclusion by referring to the rating guidelines, saying:

> *What I would like to clarify 'complex sentences' here is…what Lavender said is true from the view of grammar. But if we only identify simple and complex sentences basing on dependent clause and main clause, 'f-a-n-b-o-y-s' [coordinators] or even though passive voice and comparison more…more or many more, there is no place we can count those sentences. But it is obvious that those sentences must be* ***better than simple sentences****. Then we will identify them from different dimensions, not only based on main clause and dependent clause. Then here this TT clearly has an attempt of using sentences apart from simple sentences. (MD_Recording 2)*

This clarification from the rating guidelines appeared to echo two of the variables suggested by Norris and Ortega (2009) (see chapter 2, section 2.8.3) by counting both subordination-based and coordination-based variables. Daisy's

conclusion was found to help bring more agreement toward the raters in their identification of grammatical complexity in their TAPs, except for Tulip and Lotus. Lotus's typical comment was "can use complex structures such as 'because', 'but'" (Lotus, N, TT9), which exemplifies that the raters (including both the experienced and novices) identified the complexity of grammatical construction and accuracy in the TAPs. They identified complex structures by attending to the subordinators indicating a dependent clause (because) and the coordinators combining two independent clauses (but) as agreed in the moderation discussions. The raters also paid attention to the accuracy of the complex structures used and the range/variety of the complexity; for example, Tulip (E) frequently commented:

> *Complex structure...has "when" bu**t not really correct***
>
> *Uses conditional sentences **correctly***
>
> *This TT in part 2 performs **popular complex structures**, seems to control very well. Sentences with "when", "if", "so" she uses repetitively. Can say she uses **simple complex structures** skilfully, simple structures, uses them many times, good.*

or Lotus (N) said in her TAPs:

> *Is using complex structures, but still **simple complex structures***
>
> *This TT uses both simple and complex structures. The complex structures are **quite clear**. But the number of complex structures is **not varied** yet.*

Apart from identifying and evaluating the complexity of the structures used, Lotus and Tulip seemed to classify complex structures into two types – frequently used and less frequently used – in their attempts to evaluate the variety of complex structures (which appears in band 8 descriptors). This way of classification appeared to be slightly different from what had been discussed in the moderation discussions; however, it was evident in the TAPs data of the raters. Tulip and Lotus seemed not to completely agree with the conclusion of defining "complex structures"; thus, they developed this classification strategy

to compromise between their own definition and the rating guidelines. This seems to indicate that some raters would not be willing to give up their perceptions of the assessment criteria which was in contrast with what was required in the rating guidelines. However, this rating feature was not found in other raters' data, which may suggest that the moderation discussions were generally effective in creating more agreement among other raters in their identification of grammatical complexity.

Additionally, Lavender (N) reported in her TAPs that she found it difficult to differentiate between "some control" and "good control" of grammatical structures. This point was elaborated further by Peony (N), who said: "I do not know how many is at B2 or C1" and requested "a quantifiable measure" (interview). The interpretation of the use of abstract words such as "some" or "good" appeared to be a greater challenge among novice raters rather than experienced raters. The fact that this kind of challenge in descriptor interpretation was not mentioned by any of the experienced raters is of particular interest, and could be accounted for, to some extent, by the considerable number of ratings done and training received. Daffodil, an experienced rater, shared in her interview that the more trainings she attended, and the more ratings she performed, the more confident she was in her ratings. The confidence that Daffodil mentioned, I would argue, referred to the decreasing difficulty she encountered in interpreting and applying the descriptors in her rating process as she gained experience. This confidence can be seen in Vignette 1 when Daffodil could eloquently articulate what she attended to during her rating when asked to describe her rating process, which was rarely seen in novice raters' data. This confidence seems to indicate the impact of the community of practice through which the raters could enhance their shared understandings of the rating criteria and the norms.

Another feature of the grammatical structures that the raters paid attention to was the range of the complex structures that the TTs used. Both experienced and novice raters tended to quantify those structures as they used quantitative language such as “more”, “many” or counting words. The two examples below typify this feature.

> *About complex structures, the TT can use **more**…she can use “when”, “because”, relative clause, “if”, apart from the sentences with 2 clauses Subject-Verb, or the structure “so” she has used in part 1. (Daffodil, E, TAPs)*
>
> *This TT at first uses a lot of sentences, although it is still in part 1, she can use **many** complex structures, of course **only three** sentence types. (Daisy, N, TAPs)*

The raters seemed to find no difficulty in evaluating the range of grammatical structures; however, it was evident in the data that the quantification of complex structures contributed to the variations in the raters’ scoring decision. The variations are discussed further in the next chapter (section 5.3).

The data in this section appears to show that both the experienced and novice raters attended closely to what was included in the first assessment criterion – grammar, including accuracy, complexity and the range of the grammatical construction. All the raters were able to identify grammatical errors in the TTs’ speech. They also identified “the extent to which learners produce elaborated language” (Ellis & Barkhuizen, 2005, p. 139) by detecting subordination-based and coordination-based sentences, two of the three variables suggested by Norris and Ortega (2009) as reviewed in section 2.8.3. Another distinction between experienced and novice raters was that raters with more rating experience typically provided more comments and their comments were more specific. They were able to identify in detail specific mistakes that the TTs made, what structures the TT used and how many structures were employed. Moreover, all the raters appeared to give significant attention to grammatical accuracy. This is perhaps expected, given the major role of grammar in

predicting overall scores in earlier studies such as Lumley (2005). These findings, however, tend to provide more empirical evidence to support the claim that different attention is paid to the descriptors in one assessment criterion by the raters.

4.2.2 Vocabulary

The central sub-themes under this assessment criterion included:

- Good words/phrases (less frequently used, collocations, terminologies, idioms)
- Accuracy and Appropriacy/natural use (wrong word form/word choice, singular/plural form)
- Paraphrase, repetition
- Size of vocabulary (sufficient, a range, a wide range, etc.)

This rating behaviour is now illustrated to indicate how specifically the raters attended to these features.

It was evident in the data that all the raters paid considerable attention to identify "*less common words and idiomatic expressions"* (language in the descriptors) in the TTs' speech. They often repeated the words/phrases with a pleasant voice when they heard some "highlights" in the use of vocabulary. Moreover, all four experienced raters and several of the novice raters, including Jasmine, Lavender, Rose and Hyacinth consistently showed this behaviour. In many instances, including the following extract, Tulip listed the words used by TT4 which she thought to be "good".

> *certain situations, quite good vocabulary*
>
> *rewarding, quite good word*
>
> *mood*

the optimal

the optimal

relaxation, good vocabulary isn't it

Tulip (E) – TT4

The other novice raters showed similar behaviour, although less frequently and less specifically, for example Peony (N) commented: *"…can give quite many words such as "ultimate", "rewarding", "individual", "society""* (TT4). This type of rating behaviour was predominant compared with other aspects such as appropriateness or repetition in vocabulary rating. This tends to reflect Luoma's (2004, p. 16) idea that when raters regard "well-chosen phrases" in L2 speech as evidence for the richness of the speaker's lexicon they should reward this aspect of language performance in the assessment. Second language acquisition research has also acknowledged the importance of learners mastering prefabricated multi-word lexical chunks – fixed and semi-fixed expressions, strong collocations, pragmatic functions, idioms, etc. – (Boers et al., 2004, p. 54).

One explanation for the differences in the way experienced raters and novice raters commented on the use of good vocabulary might have been that some words/phrases were more identifiable to some raters than the others. In other words, some raters may perceive certain words/phrases as the highlights of the TTs' vocabulary use; but this may not be so in other raters' opinions. This issue was also expressed by several novice raters in the interview data. They seemed to highlight the need for more training on how to assess less common words, collocations and idioms. This extract below shows that vocabulary type and level identification was a struggle for Hyacinth, a novice rater. She said:

> *I personally see that I measure the test takers' ability to use vocabulary, I have rated a lot, but I can't have time to see* ***which level*** *the vocabulary test taker used belongs to, whether it's* ***sufficient****, whether they can produce vocabulary at both familiar and unfamiliar topics. Another point I find it difficult is that it says (in the rating scale) less common and idiomatic*

expressions, then with their rules like that, there are many opinions regarding how less common vocabulary is […]. (Hyacinth, N)

This narrative seemed to reveal that oral rating required the operation of complicated multi-tasking within time constraints, which poses a big challenge to novice raters like Hyacinth. The novice raters appeared not to be as confident as the experienced raters in dealing with this aspect. The raters with more rating experience were more confident in identifying the lexical sophistication and even knew which levels the words were at. Daffodil (E) said in Vignette 1 that "*Then for the vocabulary, I will see what words they are using, for example, words at level B1, B2 or C1, things like that can be noted all*", which was similarly shared by other experienced raters, including Orchid, Tulip and Sunflower. I would argue that this provided another piece of evidence in response to the question of how experienced raters differ from novices. It deserves a comment here as Hyacinth had fewer years of teaching, and considerably less rating experience than Daffodil, so it might have been more difficult for her to recognise the level of words than for an experienced, trained teacher-rater like Daffodil. This suggests a closer investigation into what could have helped the raters to be more confident in this aspect would be beneficial and thisis explored in detail in chapter 6.

The raters were able to identify not only the frequency of words used but also whether they were used correctly and/or appropriately as the data show. The two typical examples below show that both novice and experienced raters were able to identify the wrong forms and/or choices of words that the TTs made.

... The TT made quite many mistakes of word forms: the domestic, make you confusing, should limited. I didn't categorise them into grammar mistakes because from start to finish she used "can" properly, except "should". So I put this into vocabulary mistakes, not grammar ones. Iris (N) – MD

...um but used some words wrong, for example, "helpful things" – should be "useful things" or "she choice presents carefully", but later corrected as "choose", but should be "chooses". Orchid (E) – MD

They also paid attention to the appropriateness of the vocabulary items used in particular contexts, as illustrated by the following examples.

Sunflower (E) - TAP

> *Has used vocabulary not really accurate, right, atmosphere and teacher but used friendly and cute, these two can't use the same adjectives.*

Lotus (N) - TAP

> *Has used the word not accurate – "good-looking" (The TT used "good-looking house")*

These examples seem to indicate that the raters could provide meaningful assessment of the lexical resources of the TTs displayed during the test, which supports Schmitt's (2009) argument that correctness of use is crucial, and yet the field is still struggling to find a way of measuring such appropriacy of use in any other way than by human judgement. This contention is echoed by Shaw and Weir (2007) who, in the context of written TT outputs, noted that quantitative measures such as lexical density, lexical variation and lexical frequency profiling are not sufficiently robust to distinguish meaningfully between test takers of different levels.

Although it was evident that the two groups of raters were able to attend to both accuracy and appropriateness of the words/phrases used by the TT, the raters with more rating experience seemed to attend more to this feature than those with less rating experience. This seems to suggest that experienced raters had paid more sophisticated evaluative attention to these speech features than those with less rating experience.

Another aspect of vocabulary that received attention from the raters was whether the TT used paraphrasing to avoid repetition of vocabulary items and whether they used certain words/phrases repetitively.

Jasmine (N) – TAP

> *Vocabulary at first sounds good, for example, integrated, social activities, know how to paraphrase*
>
> *Um can use "broaden" instead of "widen"*

Tulip (E) – TAP

> *This TT's vocabulary paraphrasing is very good, good paraphrasing […]*

Another important aspect described in the rating scale was the range of vocabulary. The raters were expected to attend to this feature and evaluate whether the TT possessed "sufficient vocabulary" (Band 4) or "a range or a wide range of vocabulary" (Band 5-8) or "a good command of broad vocabulary" (Band 9-10). There was evidence showing that the raters in this study made comments on this aspect as illustrated below.

Daffodil (E) – TAP – TT1

> *The vocabulary is just at sufficient, which means being able to talk about familiar topics only. She still has the tendency of not understanding the questions, so it proves that her vocabulary size in familiar topics is not good enough to reach "a range", just at "sufficient".*

Daffodil evaluated the vocabulary size of the TTs by drawing on their ability to understand the questions and to talk about the required topics. However, these types of comments were considerably less frequent than the other aspects of vocabulary such as identification of good words/phrases or repetition. This was elaborated by Peony in her interview "[…] *because I don't know how much will be B2, C1. B1 and B2 is quite clear but B2 and C1 is my biggest concern, it's good if these can be quantified*." The raters seemed to request an alternative measurable way of assessing productive vocabulary size. The fact that this type of comment occurred considerably less frequently could be indicative of the challenge inherent in measuring productive vocabulary and/or the enormous challenges in defining the complex constructs of productive vocabulary

knowledge as discussed in chapter 2 – section 2.8.3.

4.2.3 Pronunciation

In general, the raters paid attention to all the features listed in the rating scale, as mentioned in Vignette 1; however, it seemed that they did not pay equal attention to those features. This rating behaviour is illustrated in the explanation below.

Individual sounds

The raters' comments in TAPs and moderation discussions focused on the particular sounds which Vietnamese does not have. For example, Hyacinth (N) addressed problems of pronouncing consonant clusters in her comment "abroad, abroad mispronounced "abroad" (lack of /r/ in /br/)" (TT9), or sh sound in TT14:

> *should, inaccurate pronunciation (of sh)*
>
> *she (mispronounced sh)*

The raters also identified the lack of contrast between /l/ and /n/, which is a common mistake in some regions in Vietnam. These extracts below exemplify this feature.

Orchid (E) – TAP_TT3

> *This TT mispronounced between /l/ and /n/*
>
> *ˈæk.tʃu.ə.ni (stretching voice at ni). She again keeps mispronouncing /l/ and /n/*
>
> *Daily story? (rising voice) again /ˈdeɪ.ni/*

MD_Recording 3

> ***Lavender (N)**: This TT mispronounced /l/ and /n/*
>
> ***Daffodil (E)**: Yes this TT confused /l/ and /n/*

It is worth noting here that the confusion between /l/ and /n/ is common in several areas in northern Vietnam. There exist about 58,000,000 links in the Google search engine related to the ways of correcting /l/ and /n/ in Vietnamese. This number suggests this local language feature is a common concern in Vietnam's context and it is something that the speaker is encouraged to change. In some cases, people who are unable to distinguish /l/ from /n/ are sometimes even rejected from job opportunities (Anh, 2018). In these extracts, Orchid appeared to hold a negative attitude towards this feature as she repeated the word with emphasis on the mispronounced sound. This negative attitude was found in several other raters when they recognized this feature from the TTs. The issue of how this recognition would affect the raters' score decisions will be discussed in section 5.2 in the next chapter.

It is evident in the data that both novice and experienced raters tended overwhelmingly to attend to the ending sounds of the words the TTs pronounced, for example:

Lily (N) – TAPs

Lack of ending sounds

Club (should be clubs)

many time

this TT can't pronounce the ending sound - language

this TT pronounced "dance" – fail to pronounce the ending sound

Tulip (E) – TAPs

This TT has a tendency of adding /s/ at the end?

canS asks me – this TT has a tendency of adding /s/

This type of comment was typical among all the raters. They frequently commented on whether the TTs pronounced the final sounds of the words or

not. This feature 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 many endings of words. Consequently, this common feature might have been more identifiable to the raters in the study. This, I would argue, tends to provide evidence for the claim that the attention the raters pay toward speech can be conditioned by the context in which they work and the teaching experience they have.

Stress and intonation

The raters occasionally commented on these features in the pronunciation descriptors; nevertheless, this type of comment was considerably less than those on individual sounds. The example below illustrated all of the comments that Sunflower had on stress and intonation.

Sunflower (E) – TAPs

- *Word stress…no, sentence stress…no*
- *Word stress, some have such as information…but computer …no stress*
- *The TT has quite good pronunciation with individual sounds, has stresses in sentences*
- *Pronunciation is quite okay, has stresses.*

These comments were similar to those of other raters as they tended to focus on the overall impression of the TTs' pronunciation. As the impression that one's intonation may have on the listener can be dependent on the way the listener perceives it, this seemed to allow a certain level of subjectivity or flexibility in the raters' evaluations. This issue was revealed in the example below which was extracted from the moderation discussions.

MD – Recording 1

> ***Lavender (N)**: it means when we listen, we immediately realise that the way she talks is quite easy to listen and to understand, but if we listen carefully,*

she misplaces stresses in words quite often. I remember one example, the word 'continue' which is repeated many times, means that is systematic. Plus, word stress seems to be okay but her intonation is not consistent. There are some places when she remembers she can emphasise, particularly in the last part, but the very first parts it's quite flat, sounds not pleasant, not emotional, do not show emotion when speaking.

[…]

***Jasmine (N)**: first, her individual sounds are very good, her ending sounds are good, she does not miss the ending sounds. And her voice, in terms of pronunciation is easy to listen and to understand. I think generally she has stresses, like word stress and sentence stress. About intonation, personally it's not very flat to me, still showing some effort.*

This conversation between Lavender and Jasmine exemplified two different ideas of this TT's pronunciation features, particularly intonation, among different raters. The effect that the TT's intonation made on different raters appeared to be the reason for this difference. Lavender did not consider the intonation as something pleasant to listen as it seemed not to have ups and downs whereas Jasmine perceived the opposite. This may suggest that different raters had different perceptions about intonation, which is shortly discussed further in the section below, where the raters explained what natural intonation mean to them.

Natural pronunciation and intelligibility

Apart from individual sounds and stresses (both at word and sentence level), the raters were required to evaluate if the TTs' pronunciation was natural in higher band scores (6 and above). This section will first illustrate how the raters perceived "natural" and "intelligible" as component of this and later how they felt about the related descriptor.

The raters' perceptions of "natural" and "intelligible" were revealed through their contributions to the moderation discussion, which are provided below.

MD – Recording 3

> ***Lavender (N)****: natural means* ***not fake****, common, we encounter* ***in daily situations****…it's like the way we* ***communicate daily****…that means natural*
>
> ***Rose (N)****: what is "fake"?*
>
> ***Daisy (N):*** *[…] So here we will consider "natural"* ***in words or phrases****, right? Let's take "cultural shock" as an example, we will consider how to pronounce it naturally, something like that. What do you think?*
>
> ***Lavender (N)****: means it is* ***like the expectation of the listener****, normally we expect that word to be pronounced like that*
>
> ***Lily (N)****: I think natural means when it is pronounced, we can* ***identify it immediately***
>
> ***Lavender (N)****: it's just* ***intelligible****, but it's* ***not natural***
>
> ***Tulip (E)****: true, if only heard and identified, it is* ***intelligible***
>
> ***Hyacinth (N)****: if heard and can identify, it is clear*
>
> ***Lavender (N)****: true but not yet natural*
>
> ***Daisy (N)****: in fact, "natural" is* ***quite personal, a bit subjective****. There are some people who have frequent contact with a particular group, they will consider it as natural, if they contact with a different group, they will consider this as natural. So, we agree that* ***intelligible*** *is something when we hear it we can understand it immediately, but about* ***natural, we have to see how the words or phrases are pronounced****. As before when we heard the word "activity" we saw it as not natural, so we base on that to say it is not natural.*

There seem to be two lines of argument regarding "natural" in this extract, which was found in other individual moderation discussions as well. The first line of argument, which was represented by Daffodil and Lily, in this extract and by some other raters, was that "natural" was defined by the effort that the listener needed to understand the words/phrases that the TT pronounced, either easy or difficult. On the other hand, Lavender's argument, which was then supported by Daisy, seemed to be based on a number of aspects, including (1) the effect it creates on her - the listener, and (2) the expectation of how it was normally pronounced in daily communication. To them, *natural* would mean the speaker

should create a pleasant effect on the listener. In that case, it failed to have this effect on Lavender in particular. For the second aspect, I would propose that Lavender and Daisy were unconsciously referring to the "nativeness principle" which was discussed in chapter 2 in Levis's (2006) work. It is worth noting here, regarding Lavender's extensive experience of learning and practising English, that she achieved a 9 in the IELTS overall band the first time she took the test. She also received a scholarship from the Australian Government for her Master's degree. Perhaps she had listened to numerous instances of how English words were spoken in English-speaking countries via her own experience of studying abroad, and her own experience of listening to various authentic sources such as movies, news, practice tests, etc. Thus, she would have expected that the words/phrases should be pronounced and used similarly in daily situations such that L1 English speakers could understand the speech. This is an interesting finding as it appears to unpack how the rating process was experienced by L2 English raters in their context. A brief recap is necessary here related to the context of the VSTEP. The VSTEP is a localised test but global in the sense that it is aligned to the CEFR – an international framework of standards. The issues between local dimension (L1 influence) and global dimension (natural pronunciation/intelligibility) can be seen as a tension to be dealt with by the raters in this section (more evidence of this tension is discussed in detail later in chapter 5). This finding also provides more evidence to support the claim made in chapter 2 (section 2.8.2) that more investigation into the practice of English assessment in local contexts is needed, as standard English may no longer exist among L1 raters in international contexts but may still be a common concept in L2 raters in local contexts. Moreover, the different perceptions of these key terms in pronunciation also suggests that the descriptors for pronunciation in this rating scale were multi-layered, which may allow for differences in interpretation. This reinforces the work of Isaacs (2014) who argued that

pronunciation descriptors are often too ambiguous to articulate a coherent construct. The raters in A. Brown's (2006) study also expressed more clarity in the band levels.

4.2.4 Fluency

The raters' comments on this criterion seem to focus on:

- Their overall impression
- The number of hesitations and pauses (filled and unfilled pauses)
- Error correction, false starts and repetition
- Length of utterances

The majority of the comments were on the first two features: overall impression and the number of hesitations and pauses. I illustrate each in turn with several examples.

Both novice and experienced raters seemed to hold an overall impression of the TTs' fluency as their comments were often like:

Jasmine (N) - TAPs

- *Fluency **not good***
- *Ok, only fluency is **not good***
- *Fluency is **not really good***
- *Uhm fluency is **quite good***

Sunflower (E) – TAPs

- *Fluency is **good***
- *The TT speaks **quite fluently***

The use of very informal and personal words such as "not good", "quite good", or "not really" is suggestive of something intuitive in the raters' assessment of fluency. This kind of fluency perception seems to fall into what Segalowitz (2010) categorised as "perceived fluency" as discussed in chapter 2 (section 2.8.1). It

seems that to the raters fluency was the impression "on the listener's part that the psycholinguistic processes of speech planning and speech production are functioning easily and effectively" (Lennon, 1990, p. 391). This tends to provide more evidence for the claim made in chapter 2 that holistic evaluation plays a role in the rating process even though it is not part of the rating scale. The representation of different degrees of 'good fluency' and how it helped the raters in their rating is explored further in the next section.

Second, another typical focus of the raters on fluency was the quantification of hesitations and pauses and they occasionally mentioned the length of hesitations or pauses. This is illustrated in the two examples below.

Daffodil (E) – TAPs

> *Fluency is not good, seems to have **quite many** hesitations*

Peony (N) – TAPs

> *Fluency, it's clearly that even in part 1, she evidently pauses **a lot** and the sounds like 'er','um' are too many*

This focal aspect in the raters' assessment of fluency appears to be in line with one of the findings in Bosker et al.'s (2014) study, claiming that pause frequency is likely to be a more important indicator of L2 breakdown fluency than pause length. Furthermore, the detailed discussion in chapter 2, section 2.8.1, showed findings from other studies about the importance of the location of pauses in L2 research (de Jong, 2016; Tavakoli, 2011; Tavakoli et al., 2020). There seems to be no evidence in the current study that the raters mentioned the positions of the TTs' pauses. I would propose that it might have been due to the strict time limitation they had for making their ratings; thus, it appeared to be impossible for the raters to attend to this feature. Or another alternative explanation for this could have been because the raters were not trained to pay attention to

where the TTs paused, and the meaning of the location of pauses did not appear in the rating scale.

Moreover, the raters in this study seemed to pay attention to two types of filled pauses – one with 'er', or 'um' to lengthen the answers and the other with 'like' and/or some discourse markers. These extracts below illustrate the first type of filled pauses.

Rose (N) – TAPs

- *The TT answers with a lot of 'er', 'um'*
- *Too many er*
- *Too many er um, takes too long but not yet finish one idea*

Tulip (E) – TAPs

- *This TT has quite many er um*
- *This TT only answers falteringly*
- *Too many er um*

There seemed to be an agreement among the raters that this type of filled pause did not create a positive impression on them, particularly with low proficiency level TTs. The way that the raters emphasised the words such as "a lot of", "too many", and "too long" with a longer and stronger tone in their voice appeared to indicate their tiredness, even annoyance, while giving these comments. In contrast, for higher level proficiency TTs, although the raters attended to this filled pause type, they tended to hold a more positive view of this feature, as illustrated in the comment made by Daisy below.

Daisy (N)

> *Although this TT has er, um but in my thoughts, such er um are quite natural, can't always speak throughout. She doesn't er, um for too long, just for one second or two, so I noted it down but noted as natural.*

Regarding the second type, there was one instance showing that the filled pause may create different effects on different raters. This high proficiency level TT

(TT10) used “like” as a filler in her talk, and below are the comments different raters made on this feature.

Lavender (N) - TAPs

> *Use ‘like’ as a filler, quite natural. Do not overuse the filler, can maintain the flow of speech, overall it’s okay*

Daisy (N) – TAPs

> *Listening to this TT, my first impression is her very good pronunciation. Although in fact there are some places which are not very natural, particularly she uses the word ‘like’…everything she talks she inserts ‘like’, so her speech sometimes is unnecessarily interrupted*

It can be seen from the extracts above that there existed two contrasting views on this filled pause type. One positive view was represented by Lavender, Lily, and Lotus, who considered this type of filler as a communication strategy, therefore enhancing communication. In contrast, other raters including Daisy, Rose, Orchid, Daffodil, Sunflower, and Tulip, viewed it as a breakdown of fluency. The others categorised it as a repetition of grammar structures. Additionally, instead of commenting on the locations of hesitations and pauses, the raters commented on the reasons for such hesitations and pauses of the TTs, as required in the rating scale, though not frequently.

Daffodil (E) – TAPs

> *The hesitations for grammar…**the grammatical and lexical planning is quite clear**…have the feeling that, here the TT’s ideas seem to be cut off … do not know how to express the ideas in the later part.*

Peony (N) - TAPs

> *The TT has **many** hesitations although it’s part 1, she hesitates and then **finds her own ideas***

Third, the issue of error correction, false starts and repetition was another aspect that both novice and experienced raters paid attention to. However, self-

correction was something that Daisy (N) had not considered as an important part in fluency evaluation before. She said:

> *This is the part that I often skipped, by which I mean this is something in a natural performance people can talk and that they can correct what they want to say is normal. So, when I evaluated fluency before, it was mainly about if he/she had hesitations or long pauses […]*

This idea was also shared by Lilac (N) in the moderation discussion. She said that:

> *I appreciate that the TT recognizes the error and corrects it immediately. I do not consider it as a repetition, sometimes in everyday communication we encounter that a lot, we can be aware of and can stop to self-correct to make the ideas clearer to the opposite person. So, I do not criticize the fact that the TT makes mistakes and correct the mistakes.*

From the perspective of a listener (Daisy) and a teacher (Lilac), self-correction was a positive sign of a learner learning a new language. This difference in fluency evaluation was only found in the raters with less rating experience. Perhaps it can be argued that as the raters performed more ratings, they understood how these disfluencies should be assessed, hence gaining more agreement in their ratings.

Finally, the data seemed to show that the raters were sensitive to the speed features, but not as much as they were to pauses and hesitations. The raters occasionally commented on whether TTs' speed was slow or fast. They seemed to be more impressed by the fast speed of the TTs' speech.

Together these results provide important insights into the gap identified in the literature (section 2.8.1), i.e., how fluency was perceived by L2 raters and what aspects of fluency received their attention. There were differences in the raters' perception of some particular aspects of fluency, such as the effects of filled pauses and self-correction. However, these differences were only found in the data of novice raters, regardless of their EFL teaching experience.

4.2.5 Discourse management

Daffodil described in Vignette 1 that for this rating criterion "*I'll see how they develop ideas, whether they use examples, details or not, the level of appropriateness, whether they use simple or complex connectors, how they develop ideas*", which nicely summarised what the other raters also attended to. The raters' interpretation of thematic development generally focused on:

- Relevance of ideas to the topic
- Ideas elaborating with/without explanations/examples
- Quantity and quality of ideas (lengthy, unclear, interesting, difficult to understand)

These features seemed to receive different levels of attention from different raters. For example, relevance of the ideas seemed to be of significance to seven raters including Daffodil, Orchid, Sunflower, Hyacinth, Jasmine, Lavender and Peony as the comments they made on this feature were considerably more than the other raters. The comments below made by Sunflower and Hyacinth illustrate this rating feature.

Sunflower (E)_TAPs

> *I see there are many irrelevant ideas for example when being asked what are provided in the leisure centre, she answered she likes playing sports, or when being asked if she has good relationship with teachers she said she has few friends, the two answers were irrelevant.*

Hyacinth (N)_TAPs

> *The TT is off topic for the previous part. The answer was not to the point of the question raised by the rater.*

Among the seven raters who commented on the relevance of the TTs' answers, Sunflower (E) and Daffodil (E) went one step further by providing their explanation of why they thought the answers were irrelevant.

Another feature of thematic development that the raters attended to was whether the TTs could elaborate the ideas with details/examples. The raters also commented on the number of ideas and the quality of ideas at the same time. First, they appeared to consider if the TTs were listing ideas or developing ideas with supporting details. Then they considered if the ideas were clear and sufficient in answering the questions. Raters like Orchid (E) often commented on whether the TTs encountered difficulty in developing their ideas. Second, the raters were also concerned with paragraphing. They seemed to expect/be pleased if the TT could provide a clear sense of paragraphing, for example "*this TT has a strategy to develop from a topic sentence*" (Lily - TAPs) and "*can develop under the form of one big idea and then elaborate into two or three smaller ideas*" (Orchid - TAPs) and/or "*know how to conclude*" (Rose - TAPs). They also paid attention to how the TTs developed their ideas, by "*giving examples and/or explanation*" (Lily - TAPs). Sometimes they commented on whether the ideas were "*lengthy" (Jasmine - TAPs) or "reasonable […] or quite deep*" (Lavender - TAPs).

The raters also paid attention to coherence and cohesion as their comments concentrated on:

- Examples of linking devices
- Connection of ideas

All the raters attended to the devices that the TTs used to link their ideas together by listing them out; for example, comments such as "*the TT can use firstly, secondly*" were common in the data. One explanation for the raters' confidence in listing out the linking devices is that they may be easily identifiable, salient linguistic features. These cohesive devices therefore have the advantage that they can be listed and used as positive evidence of cohesion, as argued by Kang (2005). However, some raters not only focused on these

surface features of cohesion but also paid attention to other implicit resources to evaluate the cohesion of the speech. This was echoed in Hyacinth's data when she said: "*if the TTs themselves can't use connectors but the cohesion of speaking performance is just normal, then I appreciate more than the performance using connectors in a mechanical way to connect ideas*" (interview). This seems to be concurrent with what is discussed in chapter 2, section 2.8.4 that the key aspects of coherence are discourse, which is interrelated, unified and meaningful to the listener. The reference to "*contextual properties; that is the way in which it relates to and makes sense in the situation it occurs*" (Paltridge, 2000, p. 139) was evident in the data as Orchid stated in the moderation discussion "*this speech is coherent to me*" although the TT did not use a variety of linking devices. Thus, "*the job of the rater is to make sense of the text, using whatever resources s/he has available, including a lifetime of professional experience*" (Widdowson, 1983, p. 72).

Among those who gave the most comments on discourse management were Jasmine (N), Hyacinth (N) and Sunflower (E). In the interview, Sunflower mentioned that one of her strengths in English speaking was her "*idea development and organisation and the logical and convincing ideas*" (interview). Similarly, Jasmine considered the relevance of the ideas as the most important aspect in rating VSTEP speaking performances because irrelevant speaking meant "*no effectiveness of communication*" (interview), as she reported in the interview. These narratives indicate some sign of prioritising the ultimate purpose of using language as a means of communication, which can explain why they paid the considerable attention to this criterion in their TAPs comments. The difference in the amount of attention paid to this assessment criterion seemed not to be due to the difference in the raters' rating experience, but individual perceptions of assessing the criterion.

4.2.6 Others

Very few comments were found to focus on other aspects rather than the rating criteria. For example, Orchid (E) made one comment out of 357 comments in her TAPs on interactional strategy which was not related to the scale descriptors: "*this TT has a strategy by asking 'Can you repeat?'*". Other comments by the other raters were occasionally directed to the interlocutor's manner:

> *"a new question from the examiner? Not the same with the script" (Lavender, N, TAPs, TT2)*
>
> *"the examiner sometimes corrects the TT while she is talking, causing distraction for the TT" (Lily, N, TAPs, TT2)*

or the difficulty of the test question:

> *"this question seem to be a bit difficult" (Rose, N, TAPs, TT9)*

However, there was no evidence to show that these comments were taken into score decisions.

This section analyses the way the raters attended to the TTs' performances during their rating. I would argue that the rating scale exerted a huge impact on what the participant raters attended to in their ratings as their attention was in line with what was stated in the rating scale. However, part of the rating scale generated the variations in the raters' attention to the speech features, as analysed above.

4.3 How did the raters allocate their first scores?

The previous section focused in detail on the features of speech that the raters attended to while they listened to the TTs' performances. This section examines how the raters allocated their first scores in assessing each TT, for example as described in Vignette 1 that during the TTs' preparation time, the rater looked back at the notes and considered which scores she may give to the TTs. The first

scores in this section mean the score(s) that the raters nominated initially in their TAPs.

Generally, all the raters except Tulip (E), Rose (N) and Peony (N) seemed to allocate their first scores either during 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 a single score for one criterion first, at other times they allocated several criteria at the same time, and at some other times they allocated the overall score for the performance. This feature is illustrated by three typical examples below.

Allocation of single score – Lavender, N, TAPs_TT7

computer (wrong pronunciation)

ending sound

pronunciation probably band 4 because…easy to listen to, clear pronunciation, but individual sounds particularly ending sounds not correct

Allocation of several criteria score – Lavender, N TAPs_TT11

Grammar…at least band 6. Enough to answer part 1 questions clearly and easily, express quite many ideas.

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.

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.

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.

Allocation of overall score – Lavender, N, TAPs_TT5

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

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 seems to be different from what Lumley (2005) discovered in his study of writing raters who were found to attend to each rating criterion at one time and in turn. 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 whereas writing raters may have more time to reread the scripts and to refer to the rating scales, which then explains why they could give separate attention to each of the rating criteria. Another explanation for that might be that speech may prompt the rater to consider more aspects of the rating criteria than the writing scripts, so the speaking raters can sometimes have a fuller representation of the whole performance even after a short talk.

Second, none of these examples appeared to present particular difficulties to the raters. That is, they 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 occured 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 I 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 certain band may look like at times during their rating process.

Third, the raters seemed to occasionally use their holistic evaluation as a reference when nominating their first scores. The holistic evaluation appeared to help the raters narrow the range of band scores 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. Also, this appears to provide further support to the claim made in chapter 2 about the role of holistic evaluation in the rating process.

There were two occasions when Lavender (N) and Daffodil (E) delayed their score allocation until after part 2. Let's take a closer look at Lavender's comments on one of these occasions. Apart from her typical comments on salient features of the performance, there were several comments that seem to reveal the reason why her first score allocation was delayed until after part 2 as shown as below

TAPs – TT2 – Lavender (N)

> *Why can she not answer such a simple question like this? The previous question was answered very fluently…or somebody was prompting?*
>
> *Feel like the first question was prepared…the answer for the second question is off topic, do not understand the question*
>
> *Uhm this TT at first…for the first question seems to answer confidently and fluently, but for the other questions she seems not to understand, or extremely hesitant…can't think of what to answer*

This excerpt indicates that Lavender's biggest concern was the big discrepancy in the TT's performance, that is the difference between the fluency and discourse management of the answers. She seemed to have a good impression of the performance as a whole from the first question and hoped it would continue; however, in contrast to her expectation, the rest of the TT's performance was the opposite. Similarly, Daffodil had a similar reaction to TT5 when she said "*part 1's performance does not give me much to assess so I have to wait for other parts to see what the scores are. No roughly scores can be made. Usually for other TTs after they finish part 1, I can allocate the range of scores to focus on later. But for this TT, I can't get the necessary information to locate the scores*".

This suggests that the rater may form a particular representation of the TT's speaking performance after just a few utterances. The first impression then seems to play a role in her scoring decisions since the difference in the later part caused some confusion, then a delay in her score decision. To deal with this confusion, Lavender and Daffodil kept on listening and looking for more evidence before they allocated the scores. This is suggestive of carefulness in their rating behaviour.

It seems that the raters with less rating experience found themselves puzzled by allocating scores more often than those with more rating experience. For example, Sunflower (E) appeared to be more confident in nominating a score with explanation than Hyacinth did. There were several situations in which Hyacinth (N) and Iris (N) nominated several scores for one criterion when they said: "*I was quite concerned, considering if it's a 3,4 or 5*".

Tulip (E), Rose (N) and Peony (N) were the only three raters who skipped this score nomination stage. It seemed that after they collected the evidence they needed through the TTs' speech, they could come to the justification of their final scores without nominating their first scores.

4.4 What decision-making strategies did the raters use?

Section 4.2 shows that the raters clearly paid attention to all the language features described in the rating scale. The experienced raters tended to provide longer and more specific comments on the speech than those with less rating experience. Section 4.3 exemplifies the way that the raters nominated their first score(s) and highlighted the role of the rating scale in this rating stage as well as the role of holistic evaluation. This section examines how the raters decide their final scores in order to extend our understanding of the rating process. Vignette 1 does not illustrate this step since it might be difficult to articulate the detailed

stage of finalizing the scores. Only by looking at the TAPs data set could insights of this stage be revealed. Furthermore, it is evident in the data that raters with more rating experience seemed to judge their final scores using all the assessment criteria presented in the rating scale. Raters with more rating experience seemed to be more confident in making their scoring decisions while novice raters occasionally found it difficult to decide scores.

4.4.1 Matching

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

TAPS – TT1

Daffodil (E)

> *About grammar, this TT can relatively accurately frequently use simple structure. There are errors but do not distort meaning. And clearly has attempt to use complex structures, such as because, and, when, but and it's true that. Those structures are complex structures, of course still make mistakes, but the mistakes that she is committing do not really belong to complex structures. So for this TT, grammar is a 5.*

Jasmine (N)

> *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.*

Both Daffodil and Jasmine explicitly referred to the descriptors by reading out loud key words/phrases such as "relatively, accurately, frequently, use simple structures" or "systematic basic mistakes, but can be understood". The matching strategy also involved the rejection of other possible bands. The comment below

by Daffodil illustrates the way she was trying to convince herself of the best possible score for the TT's vocabulary.

TAPs_Daffodil (E)

> *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.*

Beside mentioning both positive and negative points, Daffodil also classified the lexical errors that the TT made and tried to match with the most suitable descriptor. These examples indicate the significant role of the rating scale in the raters' final decision. This was shared by all the raters in their interviews. They considered "*the thorough understanding of the rating scale*" as the most important aspect in rating VSTEP speaking performances. Rose (N) also compared the matching process of the speech features with the rating scale as "[..] *learning mathematics, just following the right formula, then insert the formula then you will feel confident. There is no creativity here*." These narratives gave the impression that to VSTEP raters, the rating scale was an essential part in their rating process, which is in contrast to earlier findings (e.g.: Shirazi, 2012) which revealed that the raters tend to slide away from the rating scale during their rating. Although both experienced and novice raters were aware of the importance of understanding the rating scale, the raters with more rating experience seemed to be more confident when talking about, explaining and applying the rating scale in their rating process. This would seem to suggest that training programmes about how to understand the rating scale were of benefit to the raters and that one short training programme may not be sufficient for them to perform the rating job.

Moreover, in their final judgement of the scores, they seemed to consider both positive and negative factors, although Jasmine tended to pay more attention to grammatical accuracy. The difference in attention to accuracy and complexity will be discussed in more detail in the next chapter. Additionally, it can be seen from the examples that Daffodil was more specific than Jasmine in justifying her decision by pointing out the complex structures that the TTs 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 Daffodil, Tulip, Sunflower, and Orchid always verbalised their score decisions for all the assessment criteria, as they were sequenced in the rating scale, after the TTs finished their performances. The extract below is a good illustration of this rating behaviour, as it was provided by Sunflower at the end of her TAPs talk for each TT.

Sunflower_TAPs_TT2

> *Grammar, I think she can have a 5 because she can use simple structures relatively accurately, which means "tương đối đúng" (translate into Vietnamese). There are mistakes but can be understood. There are attempts to use some complex structures, for example -because, -when. So grammar I'll give a 5.*
>
> *Vocabulary, um, there is one very important part she cannot express, so it is quite difficult to rate. For what she has shown, let's see between 4 and 5. 4 means sufficient vocabulary and use repetitively, have difficulty with unfamiliar topic and many lexical errors. Let's see if she has lexical errors. Many club, go shopping, learn balance between, ok can use balance, and pass B2, best option, encourage, learn more, good idea, achieve, certificate, cook special dishes, ok nothing wrong. Then I'll give her a 5.*
>
> *Pronunciation, this aspect may be her weakest part. No ending sounds, no stress, even no word stress. So just a 4.*
>
> *Fluency, many hesitations, keep speaking but there are parts she can't keep speaking for example answer for question 2, part 1. Part 2, um similar. In short, fluency deserves a 4. Attempt extended response but short.*

> *Discourse management, there is one part in her answer in part 2 which is not really relevant, the rest is relevant. Simple connectors: first, first of all, but nothing than that. Ok a 4.*

In contrast, those with less rating experience (Iris, Peony, Lily, Lavender, Lotus, Daisy, Hyacinth) 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, Lavender 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 before moving on to other less salient features. Furthermore, there were some occasions on which the novice raters found themselves torn between two band scores and it took them a longer time to decide the scores. For instance, two long pauses of around 30 seconds were found when Iris had to make a scoring decision, 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 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 the time limit that she finally decided on a 6. This type of situation occurred in the data of Hyacinth and Peony, 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.

4.4.2 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. An explanation will be provided after the two typical examples below.

TAPs_Hyacinth (N)

> *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.*

TAPs_Lavender (N)

> *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, which they then used to justify and confirm their scores. In these instances, Hyacinth 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, Lavender 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 the 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 support the observation Lumley

(2005, p. 202) made in his study that the scale wordings did not adequately describe their attitude to the text, but if reinterpreted in this way, they were close enough for the raters to accept and use them.

4.4.3 Referencing to holistic rating

As can be recalled from chapter 2 (section 2.6), I argued that there was a gap in the literature concerning the processes of assessing speaking performances, particularly the decision-making strategies employed by raters. The studies investigating this topic seemed to create the impression that the raters used analytical rating while evaluating the speech when an analytical rating scale was given to them. This did not prove, however, to be the case in the current study. In fact, although the VSTEP rating scale is an analytical one, all the raters seemed to frequently mention the overall proficiency level of the TTs. For example,

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

This seems to indicate that the raters tried to reconvince themselves of their final decision by basing it on the overall proficiency level that they thought the TTs deserved. This kind of reference was also mentioned in the interview as one of their strategies to deal with borderline performances. Jasmine said:

> *It's like I will consider several factors to see if the TT deserves B1 or B2, or if the TT has some good points, which are a little bit stronger, I will move him/her to the upper level.*

This opinion was echoed by other raters in the interview data set when they described how they decided the scores for borderline performances. 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 not to be present in the use of analytical rating scales.

4.4.4 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 Orchid (E) 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 and in the moderation discussions as well.

MD_TT1

> *So for this criterion this TT was between 6 and 7, so according to the rating guideline, we will add and subtract the scores. If we lift the score for pronunciation, then we will lower the score for this criterion or the other way round. So if a 8 for the former then a 6 for the later or a 7 for both criteria.*

The application of this strategy appears to indicate that the 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. Although the interviews were conducted individually, the raters seemed to have an agreement that they would automatically lift the TTs' scores to higher band if their performances fell between two band scores. They persuaded themselves of this lenient tendency by focusing on the good points and explaining the good attempt that the TTs had made.

> *This TT attempted to use a range of topics and vocabulary for those topics but sometimes still misused some words, for example experiment and experience. But she **tried to use** quite many idioms and perhaps these idioms*

> *were **the result of exam preparation process**…sounded not really natural but still showed that she had **good short term memory**, had retention rate with the vocabulary she had learned…uhm **still** give her a 7. (Lavender, N)*

Perhaps from their perspectives as teachers, the raters seem to understand the difficulty that a learner of English can encounter in their context. They may also have understood how hard it is to achieve the higher band score and the significance of the scores to the TTs. This perspective may explain why the raters tended to be more lenient in their scores for those performances which were between the levels. Another reason for leniency was revealed by Peony at higher proficiency levels was that "[…] *sometimes I compare [the TTs] with myself, so when scoring B2 and C1 levels I am often confused, then I have a tendency of being lenient*."

4.4.5 Using own sense

The last strategy that the raters used in their scoring decisions was 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.

TAPs_Daffodil (E)

> *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. Daisy (N) 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. This is consistent with DeRemer's (1998) view of rating as a problem-solving process and a constructive activity, where the raters had to "interpret the language of the rating scale and then reconcile this interpretation with the specifics of the text" (p.13). In the end the raters in my study tended to rely on their intuitive sense for their final decision in these cases. This could indicate that the scale is sometimes inadequate for 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. This tension, in Lumley's (2005) work, can be described as existing between "the publicly accessible and visible scale descriptors and the raters' privately inaccessible and intuitive impression; or between the raters' need to idealise and simplify and the complexity or messiness of intuitive reaction" (p. 241).

4.5 Summary

This chapter presented the findings for the first research question which focused on describing the mental process of the raters while rating and identifying the differences in the rating process encountered by novice and experienced raters. The criterion used to classify the raters was the number of VSTEP ratings they had done. The main findings were:

- The raters all experienced three stages in their rating process: consciously paying attention to the speech features, allocating their first scores and finalising the scores using 5 different decision-making strategies.
- Experienced raters appeared to pay more sophisticated attention toward the speech features.

- The raters with more rating experience appeared to be more confident in using the five decision-making strategies than those with less rating experience.
- Novice raters tended to have more variations in their assessment perceptions and their application of the rating scale, while experienced raters tended to achieve more agreement in their ratings.

There were some surprising findings, that is, the raters relied on their holistic evaluation and their own sense while using an analytical rating scale to rate the speaking performances, even though it was evident that the rating scale had a significant impact on what the raters looked for in their rating processes. This may indicate that the scale descriptors were not able to cover the complexity of the rating and/or the tacit knowledge of the raters was an important aspect to be considered in the rating process. Moreover, not all of the measures suggested in the literature to assess L2 speaking proficiency for each rating criterion were seen to be used by the raters in this study. There might be two reasons explaining this. First, the raters were not trained to use these specific measures, such as paying attention to the positions of pauses in fluency or percentage of errors in grammar per 100 words. The second possibility might be that the cognitive load in a time constraint such as a speaking test, where the raters had to perform multiple tasks at high quality in a limited time which did not allow the raters to use all the criteria effectively.

In the next chapter, I examine factors explaining the variations identified in this chapter and the variations in the scores that the raters awarded to the TTs.

Chapter 5: Factors causing disagreement among the raters

5.1 Introduction

This chapter attempts to unpack the underlying reasons by which the raters arrived at different scores, building from the previous chapter. Chapter 4 displays the rating process experienced by the raters. It analyses in detail the common stages of rating among the raters, including (1) the aspects of speech the raters attended to, (2) the allocation of their first scores and (3) their decision-making strategies. The analysis also unveils the differences between novice raters and experienced raters. This chapter is concerned with the factors contributing to their score variations. Such examination is necessary to answer the second research question "what are the factors that cause disagreements among all the raters in making score decisions?".

This chapter, thus, examines closely the comments the raters made prior to the moment they decided the scores in order to understand what aspects the raters considered and then to identify the differences in the way they took those factors into consideration. In particular, this chapter studies the set of comments the raters made on each TT performance to further an understanding of how the raters made sense of the performance and to identify the differences in their decisions. There are two main findings which could be termed significant. First, the local and global dimension of the English language was a contributor to the differences in the raters' scores. Their reference to L1 English speakers as a norm in defining "natural pronunciation" seemed to result in differences in their scores. Additionally, the treatment of local speech features tended to play a significant role in generating such differences. These results contribute to extending the understanding of rater variability in pronunciation assessment, particularly the possible influence of raters' attitudes to pronunciation assessment, a topic which has not been widely examined. Second, the individual

orientation toward the accuracy and complexity (both lexical and syntactic) and also toward idea development appeared to be another contribution to the variations among the raters.

5.2 Issues of global and local dimension

This section focuses on unpacking how the raters perceived the role of English/Englishes in their assessment practice and provides an analysis of how such practice may have led to the difference in their scores is provided. Another factor which contributed to the score variations is revealed through the way the raters treated the identified local language features in their speaking assessment.

Vignette 2 – Peony (N)

> *In my second year at university, there was a fast-track class, the students there provided tutoring sessions in which I was enlightened about some issues such as what main ideas were, what flat intonation was like. At that time my voice was commented that its intonation was flat, the voice [the way I spoke English] didn't have anything in it. So I was wondering…, they just pointed these out to me but did not guide me how to improve these. Then it was difficult at that time, you know, there was nothing but tapes. Back then listening to tapes was the main thing we could learn from. I realized that I could speak English better via listening. I could see the connection between listening and speaking; when I pronounced correctly, I could hear it correctly. Then I learned a technique to learn listening skills and frequently I imitated what had been said in the tapes. Honestly, at that time, I did not have chance to speak to native speakers. When I graduated from university and started teaching, only when I started teaching did I meet many native speakers, and then they asked me where I learned English. I answered that I learned English via tapes, and when there was cable TV such as BBC, naturally I watched them and learned to speak like them. I was surprised when being asked like that, I only learned English in Vietnam. Now I have more experience of working, when I rate in speaking tests, if I know the level of the TT, my voice will be different. If the TT is an English-majored student, I will change my voice, talking like I am talking to foreigners, more natural, more cheerful and the pronunciation is more standard. When the sounds are not like Vietnamese style, I see it is okay. Suddenly I feel something brighter than the older generations. For example, the TT2 in the moderation discussion, she spoke exactly like the Vietnamese style, the sounds are kind of chunks, which sounded very funny.*

Vignette 2 is an extract from the interview with Peony in which she shared how she learnt English. Peony had been teaching English for 5-10 years. She occasionally rated in VSTEP speaking tests. This narrative illustrates several major issues of learning and teaching English in Vietnam. First, the teaching materials are imported from English speaking countries. Second, Vietnamese learners mainly use English in classrooms and rarely have chance to use English with people from different nationalities. Third, the concept of English standard(s) and traces of L1 exists in Vietnam's context as mentioned in chapter 1. These issues are echoed in other participants' narratives and are analysed in detail in the sections below.

5.2.1 "I am kind of non-natural person" - Raters' perceptions of English/Englishes or standard(s) vs. varieties

As mentioned in chapter 4, the term "natural pronunciation" was understood differently by different raters. Some raters considered it as "pleasant to the ears" while the others said: "it should appear as it was supposed to be in daily situations". This evidently led to a difference in their scores of pronunciation in their TAPs when they had to match the TTs' speech with the descriptors of higher bands in pronunciation – natural. The TAP data of TT4, TT10, TT13 and TT15 (these TTs were at B2 to C1 level) reveals how this difference contributed to the raters' score variations. The extract below, showing comments the raters made on TT4, illustrates this point.

TT4

> *Pronunciation is her strength. Intelligible individual sounds, clearly articulated, has sentence stress, has word stress, has appropriate intonation, and very natural, isn't it. So I will give her a 9 for her pronunciation. (Daffodil, E)*
>
> *For pronunciation it's quite clear and natural, individual sounds are articulated clearly, word stress but wrong word stress like accent examiner*

say assess but she says assess so it's wrong, and intonation not much quite flat so 7. (Jasmine, N)

Pronunciation…quite difficult to hear, quite difficult to hear clearly what she is saying, need to 'stretch my ears', not completely natural, considering between a 6 and a 7. A 6 (Rose, N)

It could be seen that this TT's pronunciation seemed to be "very natural" to Daffodil, but "quite natural" to Jasmine and "not completely natural" to Rose, leading them to give a 9, a 7 and a 6, respectively. This example is among many examples in the data to illustrate how the difference in defining "natural pronunciation" contributed to pronunciation score variations. In other words, the raters made sense of "natural pronunciation" in a different way, which led to differences in their scoring decisions.

The underlying reason for this difference in their definition of "natural" seemed to be further uncovered in the interview where they explained how "natural" was viewed and assessed in their ratings. The two comments below typically illustrate the general consensus of the majority when asked about their perceptions of "natural".

What I find difficult [to assess] is pronunciation. Sometimes I find it difficult to be natural, because I myself is not a native speaker, without a standard accent. I feel I also could only speak at that level, and it will be difficult to follow native speakers, as this is kind of customization for Vietnamese. (Rose – N, interview)

The issue of assessing if the TT is natural or not is very difficult while I am kind of non-natural person. (Lilac – N, interview)

Rose and Lilac, with more than 10 years of teaching experience in EFL, both understood "natural" through the lens of speakers for whom English is their first language. It seems to the raters that "natural" was a standard attuned to L1 English speakers, which explains why it was a challenge for them to define the term "natural". This way of perceiving "natural" tends to fall into the "nativeness

principle" which was discussed in chapter 2 – section 2.8.2. As mentioned in Vignette 2, Peony stated that:

> *When the sounds are not like Vietnamese style, I see it okay. Suddenly I feel something brighter than the older generations. For example, the TT2 in the moderation discussion, she spoke exactly like Vietnamese style, the sounds are kind of chunks, which sounded very funny." (Peony – N, interview)*

The sign of being able to reduce traces of L1 Vietnamese from the speech and sounding like an L1 English speaker was seen to be positive and favoured by the raters in their ratings. This L1-related feature of rating is further discussed later in this section.

The fact that the raters were concerned/ compared themselves to native speakers is of particular interest and could be accounted for, to some extent, by the teaching and learning practice of English in their own context. Peony's narrative in Vignette 2 seems to describe that in a context like Vietnam, where English is considered a foreign language, the use of English mainly happens in classrooms where English is taught. At the time when the study was conducted, the majority of the teaching and learning materials in this context were imported from countries where English is the first language, including England and the USA. This is further illustrated in the comments below by different raters with different teaching experience.

> *The textbooks I'm using are British English, the materials I use are still American and British. For the pronunciation itself, when teaching, I also don't force them to speak like this or like that, not focus on British English or American English. (Daffodil - E, interview)*

> *Because dictionaries do not provide us with another alternative option. They have transcription for each word, but in the end they have only two speakers, one is UK English and the other is US. (Iris - N, interview)*

It seems that the broader context in which the raters were situated had an impact on their perception of "natural", and thus also on their perception of 'standard(s)'. This is despite the fact that the concept of 'world englishes' was

occasionally mentioned by the raters at the time the study was conducted, and that the raters gave the impression that they disagreed on the relevance of 'nativeness' and were strongly oriented to intelligibility. This was illustrated when Lily said, "*the more I learn English, the more often I see that English is a global language, thus it does not matter whichever way we pronounce as long as people can understand*". However, later in the interviews, only two standards – British and American – were referred to by the participants. No other English varieties were brought up when they were asked about their teaching materials. Peony explained in Vignette 2: "*I always send them the BBC videos demonstrating how to pronounce by placing the tongue, the lips in the mouth. Later I also send them the book called Pronunciation in use*." I would argue that the broader context in which the participants were immersed somehow shaped the way the raters defined 'intelligibility' and thought of English standards. As a result of this, perhaps unconsciously, the raters applied this way of perceiving 'standards' in their ratings. This finding tends to provide more insights into the issues discussed in chapter 2 in that more research is needed to unpack whether local teacher raters are applying L1 English standards in their ratings.

Adding to the impact of the broader context of the participants, the following narratives describe more fully how their experiences of English learning were, which is similar to what Peony in Vignette 2 shared.

> *At times, I listened to the people who had American voice, which was so adorable. They sounded noble. Once I heard it, I was really struck and amazed. […] Really liked American English, tried to learn American accent but I was not persistent enough. (Lotus – N, Interview)*
>
> *During the time I studied in high school, the teachers also taught pronunciation at a relative level, then I listened to music a lot, I found the sounds I made was not too different from those of native people. Unlike some people who listen but can't match the sounds. […]* I had the feeling that my accent was influenced by American English accent*. (Daffodil - E, interview)*

> *Normally I very much prefer American English. It is easier for me to listen to and understand. Looking back at the time when I just started my undergraduate study, I was given British English to listen to. At that time, my English proficiency was still low, I wondered why it was too hard for me to understand. But when listening to the same topic but spoken by an American, it was easier for me to understand and I found it more interesting. (Hyacinth – N, interview)*

Again, British English (B-E) and American English (A-E) were the only two standards mentioned in the raters' English learning experience. Lotus and Daffodil had more than 10 years of experience of teaching EFL since their graduation. Hyacinth was considerably younger, compared with Lotus and Daffodil as she had been teaching for less than 5 years. However, they seemed to share similar learning experiences where they tended to learn from either B-E or A-E in their learning materials. Moreover, they seemed to develop their own preference towards one of the standards. This preference had an impact on the raters in different ways, i.e., it had been, to some extent, a motivation for Lotus to put her effort on learning how to achieve the A-E accent while it had assisted Hyacinth in her listening comprehension. Daffodil, who thanks her music-based learning strategies, tended to be proud that her English sounds were somehow similar to those for whom English is their first language.

As discussed in chapter 3, the participants' behaviour is a consequent product of their reaction to past experiences (Keen, 1982) and their action and individual agency are embedded in a broader social context (Levy, 2014). For example, their educational experiences with English and their English teaching experiences had contributed to shaping their present and future practices of English. Moreover, the participants were in their social context; hence, I would argue that their experiences had instilled in them a strong sense of English standards, which were attuned to either B-E or A-E.

Even though the rater training programmes had attempted to reshape their assessment practice of English, the unclear instruction (see extract by Sunflower

below) of how to assess “natural pronunciation” might have allowed flexibility among the raters to adopt their own perception of “natural” which was closely related to their perception of “standards” in their ratings.

> *In the rater training programmes, I told the trainees that whether the speech is intelligible or not depends on the rater’s experience. Some can understand but the others cannot. But in our mind, we always have to be aware that* ***it must reach the standard****. Understanding is helpful in interacting with the TTs but we must assess the mistakes.* ***Being able to understand does not mean correct****. (Sunflower -E, Interview)*

The rating guidelines seemed to give the impression that intelligibility and standard were two loosely defined terms. Which standard(s) or what variations from the standard(s) were allowed and accepted seemed not be unpacked in the training.

I argued in chapter 2 – section 2.8.2 – that the practice of assessing spoken English in a local context by its local people is a less researched topic, i.e., the local assessment practice of norms and standards of English. Generally, norms and standards have been viewed and researched from international language testing bodies who deliver their tests in various local contexts (B. H. Huang, 2013; Winke et al., 2013; Xi & Mollaun, 2011) and/or from second language acquisition to inform English teaching and learning (Bøhn & Hansen, 2017; Pinget et al., 2014; Wicaksono, 2020). These studies led me to develop an expectation that there no longer existed L1 English references in spoken assessment. However, this did not prove to be the case. In fact, the evidence suggests that the raters based their ratings on ‘standard English’, which derived from their learning experiences and their teaching materials, i.e., in this study the raters tended to refer to either British English or American English as they were the only two standards they had been exposed to in their local context. Furthermore, the participant raters’ familiarisation with international standardised tests of English can account for the way the participant raters

perceived and applied these terms. I would suggest that these together with the unclear guidelines of how to define "natural pronunciation" in the local context contributed to the variations in the raters' scores.

5.2.2 "Although it is very easy to be understood by Vietnamese people…" - Treatment of local language features in speaking assessment

Chapter 4 – section 4.2.3 – stated that the raters in the current study tended to pay considerable attention to how the TTs pronounced the ending sounds, compared with other pronunciation features mentioned in the rating scale, and the lack of contrast between /l/ and /n/ did not create a good impression on the raters. In addition, Vignette 2 further illustrates Peony's attitude towards traces of L1 Vietnamese in the TTs' speech. This section, thus, shows that these local language features were treated differently by different raters, which was another contribution to the score variations assigned by the raters. In order to further the understanding of how the treatment of local language features led to difference in scores, I drew on the TAPs data set and occasional extracts from the interview data set and the moderation discussion data set. I particularly looked at the raters' decisions on the low-proficiency performances of TT1, TT2, TT3, TT5 and TT6 (A2 – B1) as they seemed to display these features more frequently.

The quote representing the issue discussed in this section came from Jasmine who said:

> *Pronunciation, although it is quite easy to understand **with Vietnamese**, mispronouncing between /l/ and /n/, lacking the ending sounds and mispronouncing a lot, so it's about 4. (Jasmine – N, TT3)*

and Lavender while they decided a score for pronunciation of TT3's speech:

> *Pronunciation obviously lacks contrast between /l/ and /n/, but **generally it's very easy to listen to**. There are mispronunciations, especially with /s/ in the*

> *middle of the words and as an ending sound. Overall, pronunciation is clear,* ***relatively easy to listen to*** *and do not see that the TT has to put a lot of effort in pronouncing the words. And can keep talking without being hindered by pronouncing words. Word stress overall can be followed. Pronunciation, a 6. (Lavender – N, TT3)*

The two raters seemed to share similar opinions of TT3's speech. They both agreed in their score decisions that the TT lacked contrast between /l/ and /n/ and had problems with ending sounds and mispronunciation. However, the difference in their decisions seemed to lie in their attitude toward traces of L1 in the TT speech. On the one hand, Jasmine in her decisions for pronunciation of other low-proficiency TTs frequently referred to the fact that Vietnamese people would find it "*easy to understand*", but she tended to award lower scores to such performances compared with other raters. This comment is a typical comment made by Jasmine in her TAPs. I would infer that she did not have a positive attitude towards the clear traces of L1 Vietnamese in the TTs' speech. This type of attitude was echoed in Tulip's TAPs when she commented "*I listen, I cannot misunderstand but it makes me feel annoyed, as the TT's pronunciation is not standard*".

On the other hand, Lavender's decision to place the TT 2 band scores higher than Jasmine for the same speech was due to the fact that the speech was clear to her and easy to understand. I would argue that what Lavender said "*easy to listen to*" meant the ease of understanding the speech. This example typically illustrates two strands of opinion on defining how easy the speech was to understand, i.e, two different interpretations of intelligibility as discussed in chapter 2 - section 2.8.1- which then impacted their score decisions. Some raters like Jasmine or Peony seemed to have strong orientation toward intelligibility beyond the local context. This means the speech should not only be intelligible to Vietnamese people in the Vietnamese context but also intelligible to different listeners in different contexts. These raters might have aligned to international

standards. In contrast, other raters seemed to be more lenient toward assessing intelligibility in the local context where Vietnamese learners mainly use English to communicate with other Vietnamese people. This reinforces the study by Isaacs (2014) which stated that the question of "intelligible to whom" has not yet been properly resolved in the Vietnamese context in general, and in VSTEP rating in particular. Additionally, it is interesting to know that Lavender seemed to be more lenient in assessing 'intelligibility', given her extensive and impressive profile (Band 9 overall IELTS and having studied for her master's degree in Australia). The findings in this section seem to extend the understanding of rater variability in pronunciation assessment, particularly the possible influence of raters' attitudes on pronunciation assessment, a topic which I argued in chapter 2 - section 2.8.2 - has not been widely examined.

5.3 Orientation towards assessment criteria

5.3.1 Orientation towards accuracy and complexity in Grammar

As discussed in chapter 4 the raters seemed to pay considerable attention to the accuracy of grammatical structures in TTs' speech by identifying errors. They also attended to complex structures based on the identification of coordinators and subordinators. The analysis also revealed that the way raters quantified complex structures was different from rater to rater, thus possibly contributing to the score variations. In this section, I investigate more closely the decisions the raters made relating to this assessment criterion by comparing and contrasting the decisions made by all the raters on each TT performance. This section first illustrates how the raters decided the scores for grammatical features in TTs' speech and then analyses what may have caused the variations in their scores. I would argue that the different orientation toward the accuracy and the complexity of grammatical features contributed to the differences in the raters' scores.

The accuracy of grammar appeared to leave a stronger impression on some raters than others, particularly with low-level proficiency speech. The comments given by Jasmine and Sunflower illustrate two typical observations made by the raters on TT2 whose speech was at A2 level.

> *Generally this TT can only speak with limitation and link simple ideas with connectors but and that. OK then look back at the grammar, use all the simple sentences not quite right, has many basic mistakes but generally can understand what she is talking about, a 3. (Jasmine -N, TT2)*
>
> *Grammar, I think she can have a 5 as she can use simple structures relatively accurately, still has mistakes but understandable. There are attempts to use some complex structures, for example 'because', 'when', then grammar a 5. (Sunflower – E, TT2)*

In the final decision for this assessment criterion – grammar, Jasmine tended to focus more on the mistakes that the TT made. I would suggest that she was aware a moment before that the TT was able to use complex structures as she pointed out the connectors "but" and subordinator "that". However, the impression of "all simple sentences are not quite right" and "basic mistakes" seemed to explain why she was more severe in awarding the grammar score compared with several other raters including Sunflower. Sunflower, in contrast, appeared to orient her decision to the effort of the TT's using "some complex structures" before deciding the score. While more severe raters had a tendency toward punishing mistakes, more lenient raters tended to reward attempts to use complex grammatical structures. These extracts typify the raters' different orientations toward accuracy and complexity in their final decisions. It is interesting to note that these differences did not appear to be related to rating experience, teaching experience or the way the participant raters attended to the speech. The difference seems to indicate a new factor which can lead to score variations, that is, the individual orientation and appreciation of one aspect of the rating criterion. The example below provides further evidence for this feature as the comments of two experienced raters on TT3 are analysed.

> *For grammar, this TT [a long sigh], get well-paid job, want to go further study, want to go further study, mistakes. But can use 'if', 'that' quite accurately. 'I don't know what', but pronouns she uses wrong here and there, 'my' 'your' are all mixed up, only a 5, cannot get a 6. This is called systematic error, her errors in pronouns are systematic. Ok a 5. (Sunflower – E, TT3)*
>
> *This TT can use quite well the conditional sentences 'if' and with good frequency. Also there are some complex structures such as 'and', 'because' but the connectors are simple. But I can see no problems with simple structures, simple structures are okay. [reading the descriptors of band 6 and 7]. I think she has good control of complex structure, so between 6 and 7. I give her a 7. (Tulip – E, TT3)*

Both Sunflower and Tulip agreed on the accurate use of complex structures made by TT3 who was at B1 level. However, as Sunflower focused more on the wrong use of pronouns and classified it into "systematic errors", she decided on 5 as the final score. This way of assessing was echoed in Jasmine's decision as she said: "*ok, let's look at Grammar, so many basic mistakes, has attempted to use complex structures but wrong, 3 and 5 then a 4*." On the contrary, Tulip did not take this type of mistake into consideration, but attended more to the complexity, leading her to award a 7 for the TT. This orientation was seen in Daisy's, Lavender's and Hyacinth's final comments on this criterion. This is, perhaps, indicative of the fact that each speech performance would leave a different impression on different raters. At the very last moment before a final score was decided, the most salient speech feature was brought to the forefront of the decision, which was not similar among the raters. It can be seen again that the difference in orientation toward the accuracy and complexity of grammar accounted for the difference in the scores awarded.

In another situation (TT4), the raters arrived at different scoring decisions due to their difference in the evaluation of complex structures of the TTs whose level of proficiency was higher. While Rose decided a 7 because to her:

> *This TT can use quite a variety of structures, but not yet flexibly and totally accurate, not yet a range. This TT can use sentences accurately and flexibly*

and can be a good control of complex structures. She can have a 7. (Rose – N, TT4)

Tulip decided a 9 for TT4 whose level was at the borderline of B2 and C1 since:

> *Every structure is flexible, a wide range. For her sentences in her performance I can rarely see the mistakes, she rarely makes mistake, almost none. The frequency of using complex structures is high. For example, relative clause with who, that; so-clause, 'because', 'before, after', 'if', all are accurate. So I think she can have a wide range of grammatical structures. I can give her a 9. So a 9 for grammar. (Tulip – E, TT4)*

The difference in Rose's and Tulip's comments was the way they judged the range of the complex grammatical structures used by the TT. While Rose came to the decision of "not yet a range", Tulip evaluated the same speech as "a wide range". This occurred frequently in TAPs of higher proficiency TTs' ratings, including TT10, TT1 and TT15 (B2-C1). This finding would seem to follow what Inoue et al. (2021) found in IELTS raters. The raters in their study reported difficulty in evaluating the range of syntactic structures. I would argue that this difference in assessing grammar could have been the result of the abstractness of the language in the rating scale as unveiled in chapter 4 – section 4.2 – that the language in the rating scale somehow caused difficulties for the raters with less rating experience in interpreting the descriptors.

Together these three typical extracts provide more insights into how the raters arrived at their scores for different proficiency level TTs and what may have accounted for the differences in their scores. In other words, the rater's individual orientation towards either the accuracy or the complexity of grammatical structures used in TTs' speech and the abstractness of language in the rating scale seemed to contribute to the score variations. It is also interesting to note that although the literature has suggested a number of variables to measure syntactic complexity, as reviewed in section 2.8.3, the participant raters in my research project seemed to rely on the detection of coordination-based and subordination-based sentences.

5.3.2 Orientation towards accuracy and complexity in Vocabulary

Chapter 4 – section 4.2.2 – reveals that the raters seemed to pay considerable attention to the richness of TTs' vocabulary by identifying 'good' words/phrases. In addition, they also provided evaluative comments on whether these words/phrases were used appropriately. Chapter 4 also unpacks the fact that novice raters seem to struggle to identify the level of the vocabulary and how wide the range of the vocabulary was, while raters with more experience seemed to be confident in classifying the vocabulary items into levels and assessing the overall size of the vocabulary. In this section, the analysis of their judgements of the TTs' vocabulary seems to indicate that there were two emerging themes: the raters' orientation toward the richness of the TTs' vocabulary and their assessment of the vocabulary size, which tend to explain the reasons why the raters arrived at different scores in assessing vocabulary

First, the raters seemed to orient towards the size of the TTs' vocabulary. For example, before finalising the score for TT3 (B1 level), Lavender said:

> *Vocabulary can be at band 7, even though considering band 6. The difference between band 6 and band 7 is 'wide range'. This TT has an attempt to use a range of topics, and vocabulary for topics, but sometimes misuses, for example 'experiment' and 'experience'. But attempts of including many idioms and perhaps these idioms are the result of test practising. Sounds not very natural but has showed a good short term memory, has retention rate with the vocabulary items learned. So vocabulary band 7. (Lavender, N, TT3)*

At the same time, Sunflower commented:

> *Vocabulary, let's see if "a range" is achieved. 'experience' and 'experiment' are used wrong twice. [a list of the good phrases] a range of vocabulary of most topics but occasionally shows effort to avoid lexical repetition, relatively high, Okay. 'Part and parcel', 'see eye to eye', can achieve a 6 (6 minus) because she can use some phrases and the quantity of… she can use some noun phrases quite okay. [examples of the phrases] okay 6, 6 minus. (Sunflower, E, TT3)*

while Daffodil justified:

> *It can be seen that for unfamiliar topics, oh let's talk about familiar topics first, she misunderstood one question, the one about advice for someone. For other parts, she can use a range of vocabulary. About familiar topics, she can perform. She has difficulty with unfamiliar topics. For part 3, she can only list the points, not elaborating the ideas, which proves that the amount of vocabulary to elaborate ideas for part 3…she has difficulty. But she can use some impressive phrases, for example 'part and parcel', which belongs to C2. But only that one, the other ones I don't think they can* ***save*** *the number of mistakes of word form and word choice. If I give her a 6, she at least has [to perform on] most topics, which means even for unfamiliar topics. So I just give her a 5 […] (Daffodil, E, TT3)*

These are the final comments made by the raters while they were deciding the scores for vocabulary. It can be seen from the comments that Lavender, Sunflower and Daffodil all referred to the rating scale in their final decisions. There were two major aspects which were taken into consideration by the raters: how large the TT's vocabulary was and how rich the vocabulary was. For Lavender, a 7 was the first option that she thought of. It seemed that prior to this consideration, Lavender had formed a certain idea of which band score the TT might have been. What she tried to do was to justify her decision by differentiating a 7 from a 6. Her justification was strengthened by her appreciation of the idioms used by the TT. Although she accepted the fact that the usage was not authentic as it was "the result of test practicing", she seemed to reward this type of attempt. Similarly, Sunflower's comment gave the impression that she tended to reward the use of formulaic expressions by the TT, which then confirmed her decision of a 6. The difference between Lavender's and Sunflower's decisions, it could be argued, lied in the way the raters assessed vocabulary size; "a wide range" to Lavender may have meant "a range" to Sunflower and vice versa.

Daffodil, in contrast to the other two raters, provided a detailed justification of how she assessed the TT's vocabulary size by evaluating the difficulty the TT had in talking about both familiar and unfamiliar topics. Moreover, Daffodil also emphasised the accuracy of the vocabulary produced by the TT. An attempt to

use one C2-level phrase could not "save the number of the mistakes" the TT had made. Daffodil's judgement seemed to indicate that her impression on the TT's difficulty in answering the questions and the accuracy of vocabulary was stronger than the richness of the vocabulary, as she did not mention other formulaic expressions the TT attempted to use. The different focus on different aspects of the assessment criterion and the individual assessment of the vocabulary size appeared to explain the difference in the way the raters arrived at their scores.

Another interesting result presented in this section was the difference in the raters' judgement of good words/phrases, particularly their own way to resolve the question of how many good vocabulary items would be sufficient for each level of proficiency, as detailed below.

> *This TT's vocabulary is quite broad [listing the good phrases the TT used]. I can see she can have a wide range, has 'of most topics'. She can avoid lexical repetition, even with unfamiliar topics.* ***But*** *for less common words… I* ***appreciate*** *the collocations used [examples of the collocations], she can* ***touch*** *that point so I give her a 8 (Tulip, E).*
>
> *This TT's wide range is okay, the* ***only issue*** *is the less common words and idiomatic expressions. She does not seem to have this feature, so a 7 for vocabulary (Daisy, N).*
>
> *Vocabulary, this TT can use quite…the vocabulary is quite good [examples of good phrases]. Can be considered a 6. A 6 (Orchid, E)*

Tulip's and Orchid's comments reveal that they were able to similarly identify the good phrases that TT11 (B2 level) used; however, their individual evaluation of this feature had led them to arrive at different scores. All three raters appreciated the use of those good phrases, but Tulip arrived at an 8 while Daisy gave a 7 and Orchid decided on a 6. It might have been the case that Orchid would have demanded that the TT would demonstrate a higher number of those good phrases so that she could have given the TT a higher band score. The reason why Daisy arrived at a 7 not an 8 was "the TT's lack of less common and

idiomatic expressions", a descriptor in band 8. On the contrary, Tulip was content with the number of good phrases the TT could produce and Tulip seemed to classify collocations into less common expressions, which led her to decide on an 8. This way of classification was not something that Tulip appeared to be confident in doing as the use of "but" and the pause […] is indicative of the fact that she was aware of the descriptors and knew that "less common words" was one descriptor required for performance at band 8. However, Tulip still decided on an 8 due to the number of collocations that the TT used. She tried to persuade herself with the use of the word "touch" for this decision.

This difference in vocabulary evaluation was further elaborated by Orchid in her interview. She stated: "*There are some teachers, for them, while I am considering band 6 for vocabulary, many other teachers consider it as rare words and as already many. For me I don't see that many*" (Orchid, interview), Peony added:

> *"When I listen, I can pick up or make a list of words which are at C1, which words are only at B2, but for a performance, roughly how many words like that are okay or how many idiomatic expressions are okay, if this can be quantified, that would be great" (interview).*

This provides empirical evidence to support the claim made in section 2.8.3 that human raters may differ in their way of assessing productive vocabulary. The data showed that the participant raters' ways of assessing different aspects of vocabulary was not in line with the suggested variables in the literature. It is argued that this difference contributed to their score variations in assessing vocabulary.

It is worth noting that only three raters (Orchid, Hyacinth, and Peony) among 13 interviewed raters explained their concerns about assessing vocabulary when asked to rate how easy they felt it was to assign their ratings for five assessment criteria. The other raters seemed to be confident in their vocabulary assessment

as they similarly made the point that: "*for vocabulary, my job is just to note down some words*" (Sunflower – E, interview). I interpreted some words to mean the good words/phrases the TTs can use. The raters were able to identify good words/phrases. However, the problem appeared to lie in the question of how many were enough for one level of proficiency, and the data indicated that the raters had different ways of solving this issue.

5.3.3 Orientation towards descriptors in Discourse management

Discourse management was one of the assessment criteria which received most comments from the raters in the TAP data set. The analysis in chapter 4 – section 4.2.5 indicates that some raters not only focused on the surface features of cohesion (i.e., explicit use of connectors) but they also paid attention to other implicit resources to evaluate the cohesion of the speech. The tension between explicit use of discourse markers and idea elaboration seemed to explain the difference in the raters' final judgement of the scores. Moreover, their judgments also took the organisation of ideas into account. The individual preference toward a certain type of idea organisation appeared to contribute to the severity or leniency in their ratings.

One reason which could help explain why the raters' decisions were different was their individual orientation toward either the explicit use of connectors or the relevance of idea and idea elaboration. TT3's performance was a typical example to illustrate this point. Tulip decided this was a 7 since she was impressed with the variety of the linking words used by the TT even though she was aware that "*the TT's ideas were not many*" (Tulip, TAP_TT3_DM). In contrast, Jasmine who considered the relevance of the ideas as the most important aspect when she rated VSTEP speaking performances (interview) arrived at a 5 because "*many linking words but are repeated […]; ideas are elaborated but vague, some relevant, some irrelevant*" (Jasmine, TAP_TT3_DM).

Likewise, in TAPs of TT8, Daffodil awarded a 5 because "*the TT can use complex connectors quite well […], but she cannot use appropriate details and examples to elaborate ideas*" (Daffodil, TAP_TT8_DM). Lavender seemed to appreciate "*the attempt to use discourse markers so that the listener can know which idea she was at*" (Lavender, TAP_TT8_DM) before she decided a 7 for TT8. It can be seen from these examples that the individual appreciation of one aspect of the criterion contributed to the difference in the raters' scores. This has provided empirical evidence to extend the understanding of raters' score decisions when they were evaluating L2 oral performances.

Another factor which influenced the raters' decision was their attitude toward the organisation of ideas. This feature can be exemplified by the following extract in which they justified the scores for TT10.

> *Discourse management is okay despite the fact that this TT has the tendency to talk around and around, not going straight to the point to answer the questions, but eventually can manage to give answers to the questions. Can have quite many elaboration although it is possible to organise those supporting points more clearly, using the markers more obviously so that the listener can easily follow. If this was done, it would be more effective. But still can be at band 8 (Lavender, N)*
>
> *Discourse management, the TT has relevant choice, can elaborate ideas, but need to use more complex connectors, with relative ease, some of appropriate details, so only a 6 (Peony, N)*

Lavender seemed to start her justification of the score with her overall impression of the TT's discourse management by saying "discourse management is okay", then she evaluated the relevance of the ideas. She also evaluated whether the ideas were elaborated and well-organised. The last aspect Lavender evaluated was the use of markers. Peony's comment appeared to be similar to Lavender's as it also included the aspect of idea relevance, idea elaboration and connectors. Her previous comment on the TT's organisation of ideas was similar to Lavender's as she stated: "*the ideas are not well organised*". However, the

two raters arrived at totally different scores. While Lavender decided an 8, a 6 was Peony's final decision. This type of difference was evident in the TAPs data, which is an interesting finding as it is indicative of the fact that there must have been something else which impacted their decisions apart from the speech features and the descriptors. To unpack this difference, the other sets of data were helpful. The moderation discussion with Peony revealed that she seemed to have a preference of how a talk should be organised. She said:

> *For discourse markers, I do not like part 2 performance much because she does not give the direct answer. She mentions the ideas one by one, I feel that the structure is not okay, because usually I often tell my students that they directly say which option they select, then explain the reasons why, it's like an opinion essay and they will counter the other options and finally come to the conclusion. But this TT in turn mentions one by one. I do not like this structure much. (MD Recording 1, DM)*

Her preference toward the directness of the response and an opinion-essay-like structure might have made her more severe in deciding the score for this assessment criterion, compared with Lavender. This preference was mentioned several times in Peony's TAPs when she commented on how the TTs performed in this criterion. The TAPs data set also showed that Peony seemed to be consistently more severe in awarding scores in the moderation discussion since her scores were often lower than other raters'. This finding could arguably be important as it indicates that the decisions of scores seemed to rest in the personal understanding and perception of the listener on "the degree to which sentences or utterances in a discourse sequence are felt to be interrelated rather than unrelated" (Celce-Murcia et al., 1995, p. 15).

5.4 Summary

This chapter presents the findings for the second research question which focused on the factors contributing to the variations in the raters' score decisions. Throughout this chapter I have argued that the particular context that

the raters lived in had a significant impact on shaping how the raters made sense of English standards, consequently contributing to the variation of scores the raters awarded for pronunciation, one rating criterion. It was evident in the data that their experiences of learning English and teaching English as a foreign language were formed in a context where all of the learning and teaching materials were imported from English L1 speaking countries and where English was mainly used in classrooms. The lack of varied contexts where the participant raters could use English seemed to narrow their views and attitudes of English standards down to two standards, British-English and American English. Furthermore, their attitudes towards the traces of Vietnamese in spoken English performed by Vietnamese learners was another factor explaining the difference in rater severity and leniency in awarding their scores.

I have also argued that the individual orientation to and appreciation of one or more aspects of the rating criteria was another factor causing disagreement in the scores among the raters. For example, some raters tended to punish grammatical errors while others appreciated the effort of using complex grammatical structures in their final decisions of scores. There were some surprising findings, that is, the raters seemed to attend to the same speech features but arrived at different scores due to their individual appreciation of some certain aspect(s) of the assessment criteria. In addition, the literature led me to develop an expectation that raters can arrive at the same scores but for various reasons (see section 2.6). However, this seems not to be the case in my study.

To sum up, the main findings presented in this chapter were:

- The raters' attitude toward English standards and local language features appeared to influence their score decisions

- The individual orientation towards one or more aspects of the assessment criteria seemed to contribute to the variations in the raters' score decisions.

In the next chapter, I attempted to unpack how the raters developed their rating practice over time.

Chapter 6 – How raters develop their rating practice

6.1 Introduction

The purpose of this chapter is to explore in detail how the experience of the raters who rated VSTEP speaking tests developed over time. 13 of the participants were interviewed about their experiences of being a VSTEP rater. The descriptions of the experiences of the raters shed light on three major themes:

- Feeling overwhelmed at multi-tasking under time constraints
- More control of doing the tasks required
- Knowing what to do with confidence

The participants seemed to experience different stages, emotionally and strategically. While phenomenology is concerned with the person, it does not only focus on thoughts and feelings, but an embodied "being in the world" (Merleau-Ponty, 1962) as argued in chapter 3. This means that these raters, reflecting on the rating situations in which they found themselves, were responding to this context, i.e., responding to the world they live in. Thus, it is important to understand how the raters made sense of the context they were in before exploring how their experience developed. The following sections explain these themes in detail.

6.2 The context

Vignette 3 – Iris (N)

> *Before, when I came into contact with the VSTEP, when I didn't have any training, I felt it was like a jigsaw. I have been familiar with exams like IELTS, TOEFL, or FCE, CAE, then I feel that the first part [of VSTEP test] is quite similar to IELTS. The next part, which gives a problem to solve, is quite similar to one part of FCE or CAE, I don't remember very well. There are several options, whichever you choose, this is quite similar to FCE or CAE when they also have a picture/photo, here it [VSTEP speaking part 2] is in*

words. As for the last task, it is quite similar, the first impression is that it is quite similar to IELTS except for the mind-map. Altogether it feels like a jigsaw, the changes make it a little bit different.

After being trained, and understanding their [the test developers] approach, I understand that these things are just the format. The decision made was based on a theoretical background. After this exam, they [the test developers] can check the skills of the test taker, then I feel it is ok. If I let people explain it that way, I think it makes sense: what people want to test and the results they get, it also works together, it also matches.

Not all the exams have such a detailed rating scale with band descriptors as the VSTEP test, the other tests only have a few points to rate, and it's subjective. In the VSTEP test, it is inevitable that there is a subjective part from myself [as the rater], but at least there is an orientation for people to follow a certain mindset, that's what I learned. However, because of that, rating VSTEP tests is more stressful than rating in the other tests. The other exams allow me to rate holistically while VSTEP tests need to be rated analytically.

Because they take the test [VSTEP], there must have been something in their purpose so that they take the test. So when I rate, I have to try to judge them accurately according to the scales in order not to put them in a disadvantaged situation, sort of not being disadvantaged. For example, they should have achieved [their purpose], but they did not because of me.

Iris was a novice (N) rater at the time the data was generated (mid 2018). In common with the other participant raters in this research project, the decision to be a VSTEP rater was not one she made herself but was made by the faculty where she worked. Vignette 3, consisting of 4 passages extracted from Iris's interview, portrays the lived experience of Iris as a VSTEP rater. The vignette described how Iris's experience with the VSTEP test started and how her perception of the VSTEP test had changed. Iris, initially, did not seem to have a positive opinion about the VSTEP when she compared and contrasted VSTEP test tasks with other existing international tests. The use of the image "jigsaw" implied that the VSTEP was not authentic in itself, which led Iris and other raters to harbour suspicious opinions of the test (this is discussed in more detail later in this chapter). However, after more involvement with the test, Iris understood the reasons behind VSTEP's existence and developed a more positive attitude toward the test. With further involvement in VSTEP tests, as a VSTEP rater, Iris

understood the significance of her job and its potential impact on the TTs whom she rated. There were three issues which could be identified from this extract: (1) the issue of trust between a localised test and international standardised tests; (2) the understanding and appreciation of a home-grown test, and (3) the significance of being a VSTEP rater. These three issues were echoed by the other participant raters and are analysed in sections 6.2.1, 6.2.2, and 6.2.3 respectively.

According to the definition of "locally produced" tests by Dunlea (2013) and Wu (2014, 2016), the VSTEP is a home-grown product, i.e., it was developed by Vietnamese language testers to assess L1 Vietnamese candidates. The VSTEP also falls into the category of Glocal Type 2 tests as defined and categorised by Weir (2019), since it is localised in certain ways but global in others. Since the VSTEP was the first-ever standardised test in Vietnam, it is important to understand the participant raters' thoughts about its existence.

6.2.1 "A jigsaw test" – the issue of trust between a localised test and international standardised tests

The first issue emerged from the interviews with ten of the participants who chose to talk about what they thought of the VSTEP test the first time that it was introduced to them. The following extract nicely summarises their first experience with the VSTEP.

> *In fact, when it comes to exposure with the VSTEP, I was mainly exposed to the speaking and writing skills because listening and reading, to be honest, I have **never** actually written the exam papers and **rarely** had a chance to have a good look at the exam papers, probably one or 2 times, so it's very difficult to say what I think about that. However, I partly felt that, initially, I felt the VSTEP **took** one part of this exam and **took** one part of the other exams, it was built up like **compiling** them all. (Daffodil – E, interview)*

Implicit in the narrative is that VSTEP was a top-down policy. The raters were not aware of it until it came into effect. Furthermore, they did not have knowledge

of all the sections of the test. Without prior knowledge of it, the raters compared it with other existing international tests. As mentioned earlier in chapter 1, the VSTEP was the first ever standardised test of English in Vietnam. At the time when the VSTEP was released, Vietnamese learners of English were familiar with several international tests of English including IELTS, TOEFL iBT, TOEIC, etc. It could be assumed that Vietnamese learners, particularly EFL teachers, were familiar with international standards of English, as was portrayed in Vignette 3. Bearing this in mind, initially the test gave them a feeling of something not original, nor authentic, as if it were a "compilation of different tests". The use of the word "compiling" by Daffodil and "jigsaw" by Iris seems to indicate their suspicion of the test. Possibly, this suspicion resulted from the top-down policy of the test, as was elaborated further in Daisy's interview. She said:

> *In the past, I didn't understand the VSTEP test because I **didn't** participate in the item writer courses and courses related to VSTEP. Actually, the original format of the VSTEP did not reach the users. Teachers just **heard of** the description very generally, and there was no sample test, and until later I knew that the sample test was the book sold in the institution bookstore. Teachers didn't know what it was. (Daisy – N, interview)*

The lack of involvement in the test development stages and the lack of information about the test seem not to have good impact on the participants' perceptions of and attitudes towards the test. Furthermore, another reason can be unpacked from Sunflower's interview. She said: "*Before in faculties, almost all of the tests were copied versions from certain sources. When I was a student, I did the tests of the faculty, but some students had practiced the tests before and they had very high scores*." This type of testing experience might have been common in the learning experiences of other raters, and this occurred before the VSTEP test was released. This may have led the raters to have an unfavourable first impression of the VSTEP test.

The raters went on to describe their continued experience with the test. Their narratives seem to reveal that their attitudes to the test had been more positive when they became more involved with the test, i.e., participating in rater training programmes and being a rater. Daisy said:

> *So it was **not until then** [more work with the VSTEP] that I found it has different characteristics compared to other tests. There is something to match with Vietnamese students, because, in fact, Vietnamese students **often speak only one or two sentences** and **never extend their responses**. Those things [speaking tasks] will give people an opportunity to express themselves and it more suits **Vietnam's response culture**. That is my opinion. Part 3 helps students **have an idea** because **Vietnamese students' background knowledge is quite limited**, so when there are suggested ideas like that, it is a kind of stimulus suggestion so that **people are able to speak**. (Daisy – N, interview)*

It is interesting to note that in Vignette 3, VSTEP test tasks were perceived to be adaptations of tasks in other international tests (IELTS, FCE, CAE) "to make it a little bit different". However, with further involvement, the test tasks were now appreciated by the raters as being suitably localised, as they helped Vietnamese learners specifically to better demonstrate their English ability. This point was echoed in the other raters' interviews. Their attitudes started to change more positively after further involvement with the test. The involvement here refers to the training programmes the raters received, as those programmes helped the raters clarify that "*the decisions for the test tasks were based on some theoretical frameworks*" (Iris – N, Vignette 3) and how the test tasks could assess what they were intended to assess.

6.2.2 "A self-designed and self-written test" - the understanding and appreciation of a home-grown test

The second main issue that emerged from the data was the sense of the ownership/homegrown nature of the test among the raters. As illustrated in Vignette 3, Iris, after rater training, understood how the VSTEP was developed.

She referred to the use of “theoretical background” as evidence to put her trust in the test. This piece of evidence seem to prove to her and other participant raters that the VSTEP test was not like other tests she had known before, which were “cut-and-paste” tests (T. N. Q. Nguyen, 2019, p. 79). Therefore, they tended to be proud of the test and put their trust in the quality of the test. One typical example to illustrate their increased trust in the test is illustrated through comments from Sunflower.

> *I am quite impressed at the tests because all of the tests are self-designed and written. […] I think, firstly, when we write our own tests, the sources are more reliable, and are not overlapped with any other sources. Regarding test quality, it is relatively good; the tests are controlled through different rounds. (Sunflower – E)*

Sunflower seemed to link the test with the ability to develop the test. A brief note about Sunflower is necessary as she had several roles related to the VSTEP test. Not only was she a rater with extensive rating experience, but she was also a VSTEP item writer. Perhaps this insider’s experience with the VSTEP might have been the reason for her increased trust in the test. However, Rose, a novice rater without such inside experience with the VSTEP as Sunflower, shared her positive attitude of the test from the view of an experienced EFL teacher.

> *I think the VSTEP is the first standardised exam in Vietnam and has had an impact on the teaching methods and the assessment and evaluation process. Previously, assessments only were conducted within each division and each course. For example, when I taught course A, course A was taught first, but the assessment could even be more difficult than course B [which was taught later], where there was no consistence in evaluation. Now the VSTEP is the answer for some of problems regarding assessment and evaluation, as it can synchronize, examine, and assess students according to a set of criteria. (Rose -N)*

Rose, with more than 10 years of teaching experience, had witnessed the changes in the EFL curriculum and to assessment practice which were related to the VSTEP test. To her, the VSTEP seemed to have created a positive impact on the teaching and assessing of English in her faculty, as she viewed the test as a

remedy for assessment-related issues. Additionally, Rose compared being a VSTEP rater with the experience of working in a professional environment that she often saw in other international tests that she took. This seems to be a responsibility that she was proud to undertake. The image of professionally working in international tests is an atypical example, as it was not mentioned by other raters. However, it raises an interesting point that the test's impact may be visible to others in a particular way.

As can be seen from these narratives, the raters' views of the test became more positive as their involvement in the test developed. This is concurrent with the findings of other studies (Buck et al., 2010; Klenowski & Wyatt-Smith, 2012) as they confirmed the positive relationship between teachers' involvement in testing processes, and teaching and learning processes. Having these issues in mind, further insights into how the teacher-raters experienced their rating process in a glocalised test are of significance, as they can provide useful in similar contexts.

6.2.3 "The test will affect the identity of the TTs" – significance of the VSTEP rating job

The quotes representing this theme came from the words of Lilac who thought: "*the test will affect the identity of the TTs*", Rose who was sometimes concerned about the test "*because it affects their [TTs'] future and it's the final exam, the standard exit exam which decides their whole future*", and Iris in Vignette 3:

> *Because they take the test, there must have been something in their purpose so that they take the test. So when I rate, I have to try to judge them accurately according to the scales in order not to put them in a disadvantaged situation, sort of not being disadvantaged. For example, they should have achieved [their purpose], but they did not because of me (Iris)*

As the raters described their experience, they pointed out the special nature of this job by talking about how important the test was to the TTs. The raters

seemed to be grounded in their understanding of the importance of the scores they awarded and the potential influence of the scores on the TTs.

However, their experience with the ratings was not just that they were important, but also "worrying" (Hyacinth), "guilty" (Daisy), "regretful" (Rose) or "painful" (Daffodil). These words indicate that the raters experienced different states of mind in their experience of being a rater. For Hyacinth, the understanding of the significance of their rating seemed to put considerable stress on her as she said.

> *It was also very worrying at the time of rating. The worry about rating was that sometimes my ratings may have been uneven, and the TT might sue or request a review in case the difference was too big. (Hyacinth, N, interview)*

Due to the fact that the VSTEP is a high-stakes test, rating accuracy seemed to be central to the rating process. The raters seemed to perceive the accuracy of their scores in relation to the scores the TTs saw as appropriate for themselves. This suggests that the raters understand their scores as either an affirmation or denial of the TTs; they were concerned that the scores should be perceived by others the way they saw the scores themselves. When there was a disconnect. i.e., an "uneven" result, the possibility of an unshared sense of the scores emerged. The scores appeared to be co-perceivable by both the raters and the TTs.

Daisy described the situations when she felt "guilty" about herself:

> *Sometimes when I am flooded with other work and felt stressed, I can't concentrate on rating on that day. It seems I am not really sure with my scores, then I would feel guilty of myself when I get home. (Interview)*

"Guilty" is a word to describe a feeling of worry or unhappiness that someone has because they have done something wrong such as causing harm to another person. This comment was grounded in the rater's past experiences of rating the tests. Implicit in Daisy's narrative was that physical and mental fitness tended to

be a necessary condition for the raters to be confident in their ratings. This means that fatigue could have a negative influence on their rating performance. This is concurrent with the findings of Ling et al. (2014).

Rose, looking at her ratings from the nurturing aspect of a teacher, shared:

> *I will feel regretful if I just mark based on the exact band. I think it would have been better if they had invested more on grammar and vocabulary. (N, interview)*

Another way that the raters viewed their ratings was how they felt in the score consideration time. Daffodil stated:

> *it's okay for those TTs who have clear characteristics, but there are some who need to be considered. The process of consideration is very painful because when they just stand up and walk out, there will be another TT sitting in front of me already. And if I don't consider it quickly, the time span in my mind will go away very quickly. (E, interview)*

As can be seen from the narratives, the raters attached different meanings to the ratings they experienced, but they seemed to have an agreement among them, that is the pressure encountered by the raters in their rating experience. This pressure was a sign of the raters' being ethically responsible for the job they were doing.

This section sets out the particular context in which the participants lived by providing analysis of three emerging issues. First, there seemed to be an issue of trust in the participant raters about the new VSTEP tests compared with the long-established international standardised tests. The participants tended to be suspicious of the home-grown test when their involvement in the test was initiated by a top-down policy. However, with further involvement in the VSTEP test via participation in rater training and VSTEP rating, their attitude towards VSTEP tests became more positive. Another important theme which unpacks what the participants thought of their job as VSTEP raters is that the participant raters were aware of the significance of the VSTEP test and its potential impact

on the TTs. Thus, they seemed to be under pressure, encountering different emotions when doing their job. The next section analyses the stages the raters experienced in developing their rating practice.

6.3 Stages of development

Vignette 4 – Tulip (E)

> *I came there [the test site], they gave a test script with a bunch of things like a rating scale, marking papers, etc. Oh my god I was kind of terribly stressed. The first time I did that, oh my god looking at the script was horribly stressful. Oh, let me tell you about the very recent rater training I have just attended. That was the day when the final exam of the training took place. I'll tell you a story about a person who must have been absent from the session practising the test script. On the exam day when she had to act like a rater by asking the TTs in face-to-face mode and rate the TTs' performances, she was bewildered. I could see she was unable to figure out which part was for the rater and which was for the TTs and she ended up reading the script in the wrong order. I could understand how she felt at that moment. I was most scared of the script for the second part of the speaking test on my first day. When I looked at it, it was like looking at the wall, I did not know what to do, because there were so many sentences, some in bold and the others in italics. So, for that part [part 2] the candidate selected one option, later I would have to counter-argue the option, ah what we had to counter-argue was the first bullet point, right? The second bullet point, the third bullet point, looking at that was overwhelming, I didn't understand anything. Then after that came the words in bold and italic, then "I agree with you, to sum up, etc.", oh my god, in general I didn't understand anything.*
>
> *Unable to understand the format, let alone really bewildered, and had to tick the boxes, of course there were only few boxes to tick, now I find it so simple. But in the past, I didn't have enough time to tick. I couldn't keep up with ticking, and I had to write down a lot on a piece of paper. Then I sat until the end of the session to insert scores, to get things done. Oh my god it was extremely stressful. To be honest, in those first days, my rating was probably based on intuition. Now thinking back I'm sure my rating was intuitive. I had a bunch of things that I still couldn't understand. Having said that… those things [test scripts and the rating scale] were not public, which meant I could not keep them for myself to study them in more detail. Now we are at the training, and at least have something [test scripts and rating scale] at home to look at and think back. In the beginning, I couldn't evaluate everything, whether simple sentence was good or complex sentence was good, or how the stress and intonation was. No, not at all. So up to now I have to say that I have been better since I joined the training last summer, together with doing this for you, I find it completely different. Yes, I mean I suddenly get a lot more mature, and I have to listen more carefully. Moreover, now that I get*

> *more familiar with the documents and procedure, already memorise, my mind is no longer busy then I can focus more on their [the TTs] speaking performance. But in the past, generally I didn't catch anything at all. Vocabulary or idiomatic or something, oh my god, it was generally very difficult, it was like that in the beginning.*

Vignette 4 is part of a story that Tulip, an experienced rater, reflected on the first time she performed her live rating of VSTEP speaking performances. Similar to the other raters, Tulip quickly became a VSTEP rater due to the demand of the institution. The vignette vividly portrays the first stage of being a VSTEP speaking rater. Tulip appeared to be in a panic at her first ratings since she was not able to manage the multi-tasking, including following the test script, paying attention to the speech features and evaluating according to the rating criteria. The live rating appeared to be highly complicated and the training that she had did not prepare her sufficiently to be confident in the job at first. The repeated use of the exclamative phrase "oh my god" is indicative of the overwhelming feeling and the pressure she felt in her first live ratings. Tulip also emphasised the positive impact of the training she attended and the number of ratings she had done. These two factors seemed to help her to be better at rating VSTEP speaking performances, from her perspective.

6.3.1 Feeling overwhelmed at managing multi-tasks in a time constraint

The collective narratives shared by ten raters out of 13 raters (except Jasmine, Lilac, and Iris) seem to indicate that rating was a highly complicated and multi-faceted job and in their first ratings the raters struggled with the tension of performing these tasks under time constraints and maintaining rating quality. This was a significant theme identified focusing on the stages of rating practice development. 10 of the participants shared similar narratives of how they felt and what they did in their very first live ratings.

"I couldn't keep up with ticking the boxes. I was flooded." / Struggling with managing different tasks

This narrative shared by Hyacinth seems to further illustrate the rating procedure that the raters encountered in their first ratings and the feelings they had in their first ratings.

> *When in the exam room, I was a bit rattled at the first pair of test takers. For example, when I just turned this recorder on, the TT already started to walk in. I had already read the script, but never rated before, so I had to look at the script again, and asking the TT also made me very nervous. Because it was the first time asking, I also had to search little by little to see what the script had, then while asking I could find something different. I remember the first time I was also a bit rattled, so I was slow in asking as well. At first, I let the TT talk, but some of them talked for quite long, developing their ideas well, so the important thing I forgot to do at that time was to manage time. (Hyacinth)*

The impression left after Hyacinth's experience of her very first live ratings appeared to be anxiety, which is similar to Tulip in Vignette 4. Even though she knew what the script was about prior to the rating, she found herself not confident in using the script and she seemed to be drowning in the interlocution procedure to such an extent that she forgot to manage time for each part of the speaking test. By referring to what had been learned from the training, Hyacinth attempted to explain the challenging aspects of putting the lessons learned into practice. This seems to suggest that their understanding of the script and the rating process gained from the training did not prepare the raters enough to perform all of the required tasks smoothly when they started their rating job. This seems to show a gap between training and practice.

Tulip repeatedly described her overwhelming feelings, this time she focused on the extent to which she could not manage the seemingly simple tasks, i.e., ticking the boxes in the rating papers.

> *"I didn't have enough time to tick the boxes. I couldn't keep up with ticking, I had to take a lot of notes. And I had to sit until the end of the rating session to insert scores to get things done. […] that at first, only 4 or 5 ticks but I was completely flooded. I had to write them down on another piece of paper then later I copied them in [the rating paper]…yeah… and I even didn't know what to read and what not to read yet in the script, then the rating scale" (Tulip)*

The narrative also implied that the matching strategy at first required more effort from the rater in live ratings. Tulip had to take notes of the TT's speech so that she could match them with the rating scale to arrive at the scores. Together with other tasks including following the script, this seemed to be too much for them to handle at first. Particularly it might have been due to the time constraints and the nature of happening once only that they tried to keep the procedure running smoothly otherwise they would have been "flooded".

Sharing similar ideas with Tulip, Lily, who was novice in both rating and teaching, expressed her stressful feelings in her first live ratings due to the dual actions of making sense of the TTs' speech and the rating scale:

> *I found myself quite stressed because I had to listen to what the TTs were talking and at the same time looking at which band scores they could get. So it was really stressful, I was both straining my ears to hear and straining my eyes to see. Sometimes I had to take notes. In the first few times, it seemed that I took more notes because I was afraid that I would forget, not sure if I gave the accurate scores. (Lily, interview)*

Matching the speech features with the equivalent "band scores" seemed to make Lily tense. Perhaps, the rater did not have a chance to listen to the performance twice; she took "more notes" in case she "would forget". I would suggest that the rater tried to hold on to the TTs' speech features so that she could match them with the descriptors in the rating scale for her evaluation of the speech. It seemed to be a challenge for novice raters to retain the speech features in their short-term memory and place them in the equivalent scale descriptors. In light of this, the procedure of listening to the TT, "catching" the speech features and matching them with the equivalent descriptors happened

so fast that the raters appeared to be overwhelmed with the tasks they had to do in their rating process.

"I was stunned to see the rating scale" / Struggling with the detailedness of the rating scale

The data also seemed to reveal that understanding the rating scale and applying the criteria in their rating process was one of the tasks most raters struggled with. Rose elaborated on this tension further by describing her first impression of the rating scale.

> *"I was stunned to see the rating scale for the first time, wondering why it was so detailed and specific like that. In order to be able to score, I had to be able to remember 'attempt complex sentences' deserves band 5, must remember 5 assessment criteria, each of which was from 0 to 10, which frightened me, not to mention the academic words used in it." (Rose, interview)*

This narrative seems to indicate that working with an analytic rating scale was another challenge to raters beginning to rate the VSTEP tests. It was important to the raters that they should remember several key points in the rating scale in order to facilitate their rating process, which was hard for them at first. The mentioning of this difficulty also tends to indicate that the raters understood the significance of the rating scale in their rating processes. However, the detailedness of the rating scale with a number of key terms seemed to be a challenge to the raters rather than being helpful to them. Daffodil emphasised the difficulty she had while using the rating scale initially:

> *"I was not familiar with that at first, so I was always in the situations that required me to work continuously. Therefore, sometimes I did not even have time, both listening to the TTs and gradually scanning for key words." (Daffodil, interview)*

Iris shared similar ideas to Daffodil when reflecting on the first times she worked with the VSTEP rating scale.

> *Other exams allow me to rate according to holistic style, while rating VSTEP must be in analytic style. It is beneficial that it helps me, it forces me to pay attention to detail, but it also has one side because I often pay attention to detail, sometimes I have to look at the rating scale, I can lose tracks of what the TT is talking about if I pay attention to that rating scale for a long time (Iris)*

Clearly, for Iris, the tension between paying attention to the details of a speech and matching them with the detailed descriptors in the rating scale was a matter of both time-related and cognitive load-related challenge. Altogether the raters told similar stories of how overwhelming and daunting it was in their first ratings due to the number of highly complicated tasks they were required to perform and the pressure of maintaining the accuracy of the ratings.

At this initial stage of their rating practice, it seemed that the priority was placed on the former rather than the later, as admitted in Vignette 4: "*In the past, generally I didn't catch anything at all, vocabulary or idiomatic or something, oh my god, it was generally very difficult, it was like that in the beginning*" (Tulip).

It seemed that when the raters were busy with the tasks in their rating procedure, they were not able to pay as much attention to the speech as they reflected on their initial experience. Despite lessons learned in the training, the raters were confronted with several difficulties such as following the script, retaining speech features in their short-term memory, and matching the speech features with the descriptors.

6.3.2 More control of doing the tasks required

After feeling overwhelmed with managing different tasks at the same time, it seemed that all the raters came to the stage of feeling more in control. This theme was divided into 3 subthemes: having a plan of what to do, using the rating scale, and using different strategies to allocate the first scores.

"I focused more on planning" / Having a plan of what to do

Confronted with several difficulties at the beginning of their rating practice, the raters developed by moving forward intentionally. For several raters (Tulip, Daisy, Hyacinth, Lily), the first change they made to facilitate their ratings was to better control the test script so that they could focus more on the speaking performances.

> *After that, at other times I was more proactive. After the first times I started to know the sequence of the test, clearly know the order. The script was also no longer new to me, so I focused more on planning when the TTs signed in the list, doing those procedures, then when I started asking questions, I could adjust the time accordingly. Even if the TT said longer, I could still stop them. Generally, after the first few times, I could be able to adjust, I adjusted more regarding the time. (Hyacinth)*

The raters seemed to construct their way of rating in a more conscious and purposeful manner; they took planned steps in rating in order to achieve their goal – "*to be able to focus more on the speaking performances and the band scores*". This sense of self-control was seen through their purposeful manner of conducting the test, managing the time and interrupting the TTs if needed. This intentional adaptation appeared to be a result of the raters' reflecting on their first experience of rating. Moreover, the mention of their confidence to take the floor away from the TTs seemed to indicate that the raters understood their rating process as social engagement between themselves and the TTs. This engagement required the sensibility of the raters to find a good time to "*stop the TTs, particularly the ones who speak slowly*" as "*sometimes I want to stop them but they add one more sentence, I feel hesitant to stop them*" (Sunflower). The raters were put in a moral quandary of "*making sure that the TTs have enough time to cover all the parts [of the speaking test]*" and breaking the politeness code of social communication. The balance between strict structure and their sensibility as a conversation partner in the flow of the speaking test was a question that frequently arose in the rating process. This balance seems to

reflect the issue of the co-constructed responsibility for interactional patterns that interlocutors orient toward (L. May, 2009). As implicit in the narratives, this type of tension was not an easy task to solve for the raters who had just completed the training programmes and started their rating jobs. This finding was consistent with what Lim (2011) found in his study that novice raters seemed not to be ready to rate after a few scoring sessions. However, the findings in this study tend to provide more insights into why and the extent to which novice raters may not be ready to perform their job.

"I group them and underline key points" / Use of rating scales (underlying key words to differentiate band scores)

All of the raters seemed to move to the stage where they tried to work with the hard copy of the rating scale prior to the ratings so that it would be easier for them to refer to it during the rating process. The following extract from Hyacinth typified the way that the raters worked before they started their rating process.

> *Regarding the rating scale, normally I follow groups of competencies, B1, B2, C1, VSTEP 3-5 will have those competency groups. I group them, underline the different points in that band score. For example, where is difference between 4-5. (Hyacinth)*

This extract seems to indicate that it was necessary for the raters to highlight the differences among band scores in the rating scale because "*there are so many words and it is very detailed*" (Rose). This again provides further evidence of the fact that the raters understood the significant role of the rating scale in their ratings. They were proactive in finding a way to be in a better position to use the scale. The categorisation of different proficiency levels seems to suggest that the raters did not only consider each band score, but also took an overall evaluation of proficiency into consideration in their ratings.

> *Every time I received the rating scale, I always underlined a lot in there. After that, I gradually followed the key words to see if the scores can go up or*

> *down. But I still felt quite ambiguous because actually the meaning of words, like how is "good", how is "sufficient", how is "a range", etc., in fact, was not clarified from the first phase. So I totally felt yes, this student already had a range, but I couldn't define whether my "a range" was actually true or not true. (Daffodil)*

However, at this stage, the raters appeared to have an unclear understanding of the terminologies such as "good" or "a range" in the rating scale and needed to frequently refer to the rating scale for rating criteria and decisions. Iris said: "*I have worked with the rating scale for about 15 times, but I think I need more times to work with it to be less dependent on it*." This seems to support what Esfandiari and Noor (2018) discovered in their study that generally, novice raters tend to refer to rating scales and rely on the criteria listed in the scales more frequently when making their scoring decisions.

"Usually the starting points will be the most important points"/ Using different strategies to allocate first scores

For some raters (such as Lily) in order to deal with the overwhelming nature of managing different tasks at the same time as maintaining a focus on evaluating, they tried to refer to their initial holistic evaluation of the speech as a starting point to narrow down the area they had to look at in the rating scale.

> *First, I have a kind of impression on the TT, I'll give around this level of score. Then keeping listening, if that TT's speaking gets better, his/her score will increase. But if his/her performance gets worse, which means they are not familiar with the topic, the score can be reduced or fluctuated within a certain range. (Lily)*

Several other raters (Iris, Peony, Jasmine, Daffodil) tried to decide score(s) for the criterion/criteria they were most comfortable with as a starting point.

> *In my first times of rating, pronunciation was the part that I rated fastest because it only took 1 or 2 minutes to recognize their range of band score, and started rating from that score first. (Daffodil)*

The raters took the initiative to handle their overwhelming feelings and to better control the tasks, particularly the evaluating task. It appeared to be important to the raters to arrive at the first evaluating points/scores. However, the use of different strategies by different raters seems to suggest that the raters were not uniform in allocating these initial scores.

For the raters who had rating experience in other contexts, their ratings seemed to be situated in reference to their previous experience of rating. This reference seemed to play a role in increasing their confidence of deciding the scores for VSTEP speaking tests. It was understood as a way of confirming that the scores they were deciding were right, as Daffodil said:

> *Then I still had to be convertible because if the points of band B1 in VSTEP test were around this range, they corresponded to the points of IELTS at the same range. So I usually had to stick to such scale [IELTS]. But if I was left alone, I probably could have given up. (Daffodil)*

Sunflower explained what she had done when she was unsure about her scores.

> *If I was not sure, I often based on the TTs' (proficiency) level, anticipating their level. I would imagine the TTs at B1 would be like this like that, at B2 would be like this like that, considering all the expectation I often had towards students at those particular levels to identify where the TTs might be at. (Sunflower)*

The rating practice at this stage seemed to involve more than the knowledge and skills gained in the training as it involved holistic evaluation, the comfort they had about an assessment criterion, their teaching experience and/or their prior experience of rating in other contexts.

The change in the raters' rating practice takes place by reflecting on their lived experience and by applying a new way of thinking to their ratings. This reflection on experience enables them to realise what they acquired, what they learned, how they changed, to know themselves better and to take adequate decisions in order to provide successful ratings.

6.3.3 Knowing what to do with confidence

The actual rating and the rating volume seemed to provide the raters with an opportunity to refine and clarify their rating processes.

"I watch my watch like a hawk"

Among 13 interviewed participants, only 6 raters (Sunflower, Daffodil, Tulip, Orchid, Rose and Daisy) described how strict they were with the timing of the test. Tulip described her time management as "I watch my watch as a hawk". The 6 raters provided a detailed account of how frequently they looked at their watch and strictly set the time for each part. They also stated: "*a watch was my must-have item in the test*". For these 6 raters, timing seemed to stand out in their narratives when they were asked to talk about their current rating process. Daisy and Tulip also shared their sympathy towards beginning raters who were so overwhelmed by the tasks that they could not manage the time properly. Rose explained the reasons for her strict timing manner that "*10 minutes of performing a test is short but for TTs they want to have the chance to display all of what they have. We need to respect that.*" This seems to indicate that time management was understood as the raters' top priority responsibility to assure that the TTs had their opportunity to best demonstrate their ability. Time management was seen in relation to the effort of maximising the performance time for the TTs. This indicates not only a responsibility to perform the ratings as a duty, but also shows the caring aspect and thoughtfulness of the raters.

"I know what to look for in a speaking performance"

Five raters, including Daffodil, Tulip, Orchid, Sunflower and Daisy, seemed to reveal that they encountered another stage of rating when they showed great confidence in using the rating scale and confidence in their rating process. Daffodil, Tulip, Orchid and Sunflower were those who had the most rating

experience compared with the other raters in the group, while Daisy had started her job, which mainly involved language testing and assessment, one year before the time of the interview. Perhaps, due to the nature of her job, she had considerable opportunities to work with the rating scale. The following quotation from Daffodil typified this rating stage amongst the five raters.

> *Generally, when the TTs start speaking, I won't look at rating scale, but I'll take notes of the grammar first. I'll see if the TTs use complex structures or simple structures, how accurate they are, so normally just use the +/- sign to see in terms of quantity. Then for the vocabulary, see what words they have to search for, for example, words at level B1… First, I'll take notes… start to take into account the fact that they have one minute to prepare, then I'll see with that level, which band they are roughly at. (Daffodil)*

The raters appeared to be physically independent from the rating scale as Daffodil said: "*I won't look at the rating scale*". This showed considerable confidence in using the rating scale since in the initial stage, the raters were too "stunned" by the rating scale to know what to look at. Later, they moved on to the second stage where they rushed to highlight the key words in the rating scale so that it would be easier for them to use in their rating process. However, in this third stage, as Sunflower said: "*I have worked with it [the rating scale] a lot, knowing what is where*", so "*I do not need to read the rating scale*" prior to the ratings. This was echoed by Tulip, Orchid and Daisy. Knowing the rating scale inside and out seemed to increase the raters' confidence of using it in their rating process. This confidence tended to be the result of both rating experience and the amount of training the raters received, as stated by Daffodil:

> *…I have already had a long time doing this rating job. This is the fourth time I have participated in the training, I'm much more confident now. Because now I start to take notes, start to see what a range is, what a good range is, whether it has variety or not, or the corresponding parts (Daffodil)*

This observation was somewhat different from what Lim (2011) unpacked in his study which reported that frequency or volume of rating done contributed to positively affecting rating performance. The narrative acknowledged the

influential impact of the amount of training on the raters' development of their rating practice. In addition, this finding provides empirical evidence to support the impact of the CoP on rating practice development, as Orchid, an experienced rater, reflected on her experience of learning to be a rater:

> *For example, as I have mentioned earlier, in the past I just saw this performance was done well in terms of vocabulary, then grammar, and later I gave him/her a high score. But later on, I learned that I should have evaluated only the words relevant to the test questions, then I learned from the experience like that. That is the experience I learn from having performed a number of ratings… that might have changed me.*

For Orchid, her continued participation in the rating community appeared to help refine her own learning through interaction with other members. Tulip, in Vignette 4, also mentioned the same point when she said: "*So up to now I have to say that I have also been better since I joined training last summer, together with doing this for you, I find it completely different. Yes, I mean I suddenly get a lot more mature, and I have to listen more carefully*". This is concurrent with the notion of CoP discussed in section 2.7, that learning in a CoP is a dynamic and developmental social activity grounded in situated practice (Herbert et al., 2014). Rose, a novice rater, highlighted the significance of training for new members:

> *Obviously there must be training. You must consider it as a job which you get paid, so it must be done properly and seriously. Everyone has a different mindset, but must have the same thoughts, then it will be easier to come to a mutual agreement. It's true that we need training in order to have the same mindset.*

Rose seemed to refer to training as a way of bringing new individual raters together as it helped them to develop understanding of the norms that the community of raters held. This shared mindset was also mentioned in Vignette 3 by Iris, a novice rater, when she described how her attitude to the VSTEP had changed from negative to positive.

The second characteristic identified in the third stage was the selective notes the raters took during their rating process. Their notes were organised according to the descriptors in the rating scale, as evidence for their matching strategy later in the process. This seems to further support the finding revealed in chapter 4 that the raters' attention to speech features was heavily influenced by the descriptors of the rating scale. Moreover, retaining the speech features was important to the raters. Orchid explained how important the notes were to her:

> *I have to write them [the speech features] down in my rating process. It is because when I first hear it, I know [it's] okay, I also tick the band with my score 6 for example. However, after a while when I keep listening and I may forget, I don't know what structures they have added, for example. So I still have to note a few things to make it more accurate in my process.*

The third characteristic in the third stage was that the four raters appeared to be more uniform in the way they allocated their first scores as reflected in their interviews. All five of them used their notes of the speech features to match with the descriptors in the rating scale during the TT's preparation time. Sunflower said: "*when the TT has 1 minute to prepare for part 2 and part 3 of the test, I will read the rating scale and write on my rating paper*". It is interesting to note that in the previous stage the raters had a variety of ways to arrive at their initial scores (e.g.: using holistic evaluation, referring to previous rating experience in other contexts, etc.) at different times during the TTs' talk. However, the raters seemed to base their initial score considerations on their notes and the rating scale after the TTs finished each part of the test. The characteristic of withholding premature judgements in order to glean more information is concurrent with what Barkaoui (2010) and Wofle (1997) found in their studies.

The stories of the experience of rating (Orchid, Sunflower and Daisy), when the raters could discuss their ratings with other raters and colleagues, and the training (Daffodil and Tulip), when the raters could better their understanding of

the rating scales, contributed to the raters' increased confidence in their rating process. This appears to be consistent with the findings of Davis (2016) which stated that rater experience and training were combined contributions to scoring performance.

6.4 Summary

This chapter presents the findings for the third research question which focused on the way the raters developed their rating practice. The chapter starts by setting out the context in which the raters started their experience of being VSTEP raters. Their attitude towards the VSTEP changed from negative to positive due to their further involvement in the test (the training they received and the rating they performed). The analysis of the data also revealed that the raters were aware of the significance of their rating and this understanding seemed to be a pressure that the raters had to deal with during their development process. The raters seemed to experience three stages in developing their rating practice. First, the raters did not appear to be ready in their first live ratings, even though they were trained. After a few ratings, the raters started to learn how to manage the rating tasks more smoothly by having a plan of what to do as an interlocutor, having a strategy of using the rating scale more effectively and allocating initial scores. With further involvement in the community of raters by attending more training and performing more ratings, the raters could do their job with confidence. One of the most interesting findings was that the understanding of the script and the rating process gained from the training did not prepare the raters well enough to perform all of the required tasks smoothly when they started their rating job. To sum up, the main findings presented in this chapter were:

- The raters experienced three stages in developing their rating practice.

- The community of practice, including training and VSTEP rating experience played a significant role in increasing raters' confidence in performing their job.

The next chapter discusses the study's findings, their implications and limitations, suggests ideas for further research, and evaluates their contributions to the field.

Chapter 7 - Discussion and Conclusion

7.1 Introduction

The chapter includes a brief recap of the complete research study and provides further discussion of its findings. In particular, the chapter summarises the research aims and design in connection with the key findings according to the research questions. It also discusses the study's contribution to knowledge, implications, limitations, and suggestions for further research.

This qualitative study has investigated the rating processes experienced by the raters who rated speaking performances in the VSTEP test, a high-stakes test in Vietnam, in order to unpack the factors which influence their score decisions. The study also examined the way the raters developed their rating practice over time. Thus, the study attempted to provide rich and contextualised descriptions in order to present an authentic picture of the lived experience of the raters in performing their rating job. A phenomenological research strategy was used to conduct the study. I generated the data from three sources: observation of moderation discussions, TAPs of the raters' ratings and individual semi-structured interviews (see section 3.7). The participants were 14 VSTEP raters who successfully completed VSTEP rater training programmes, achieved C1 English proficiency level and rated VSTEP speaking tests for at least one year or more. Interpretative phenomenological analysis was used to analyse the data to answer the research questions.

Retrieving the research questions from chapter one, the next section of this chapter discusses the main findings according to the research questions.

7.2 Study findings

The research questions are listed below:

RQ1. What are the mental processes of rating speaking performances?

- What differences between experienced and novice raters, if any, are seen in attention paid to language features described in the rating scale?
- What differences between experienced and novice raters, if any, are seen in decision-making strategies?
- What differences between experienced and novice raters, if any, are seen in attention paid to test-taker proficiency levels?

RQ2. What are the factors that cause disagreements among all the raters in making score decisions?

- To what extent and in what way do these factors affect raters' decisions in their ratings?

RQ3. In what ways do raters develop their rating practice?

RQ1 was created to explore the mental process the participants went through when rating. This established a context through which to interpret the results to RQ2. RQ3 shifted the focus from the participants' rating processes to the evolution of their rating practice over time.

7.2.1 What is the mental process of rating speaking performances?

The key finding for this question was the detailed description of the mental processes that the raters experienced while rating VSTEP speaking performances. The process included three main stages. The raters, first, consciously paid attention to language features specified in the rating scale (see section 4.2) in order to allocate their initial scores for the TTs (see section 4.3). To finalise the scores, the raters used five different ways of deciding the scores: matching, simplifying key terms in the descriptors, referencing to holistic rating,

compensating and using their own sense (see section 4.4). Figure 7.1 overleaf portrays the mental processes experienced by the participant raters.

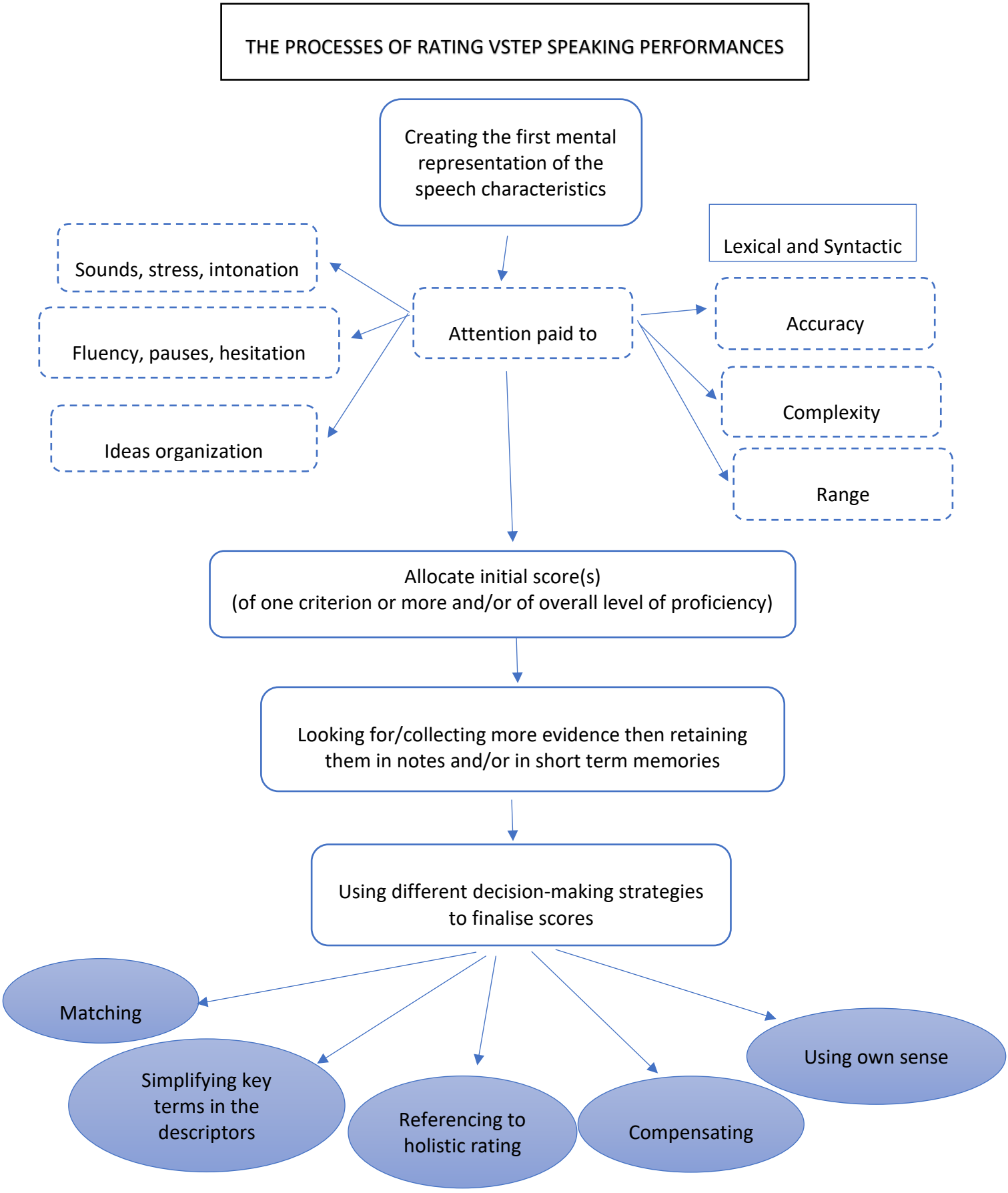


Figure 7. 1: The mental processes of rating speaking performances

This finding is important as it contributes new knowledge to the body of literature by revealing the underlying processes and the strategies used while speaking raters are "attempting to understand response input, formulate a mental representation with that in the rating scales, and evaluate the response in those terms" (Purpura, 2013, p.18). Very little was found in the literature on the question of what the mental processes of rating speaking or the rating sequence was like. Most of the rating processes and strategies are proposed through investigation in rating writing (e.g.: Cumming, Kantor & Power, 2002; Baker, 2012; Lumley, 2005) (see section 2.3). This finding 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. The finding 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.

This study also unpacks how the raters approached and used an analytic rating scale in their rating. These aspects, as argued in section 2.3, have been researched extensively in writing, but not in speaking. The general expectation in rating seems to be that the rating scale is the primary influence in rating, and that the raters have to match the text/speech to it; I agree with this. However, it is not as simple as that. This study has shown that in order to use the analytic rating scale consistently and confidently, the raters needed training to understand the scale and needed to have worked on it a number of times (at least 15 times as suggested by Iris). This seems to support the findings of the recent study (Lamprianou et al., 2020) investigating the impact of CoP on rating behaviour. Lamprianou et al. (2020) suggested that accumulating experience in other language exams was not transferable to one particular exam. In other words, the experience of working with one rating scale does not necessarily have a transferable effect to another rating scale in another exam. The difficulty the raters have in working with the rating scales (Huhta et al., 2014), therefore, should be taken into serious consideration in studies investigating rater behaviour. Results from studies which recruit lay people and offer a few training programmes to work with a brand-new rating scale should be interpreted with much caution.

Additionally, the finding that was beyond expectations was the use of holistic rating by the raters, 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. Daisy, Daffodil, Jasmine, Lavender, Orchid and Sunflower

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, which was more evidently seen in higher level proficiency TTs. 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 five decision-making strategies than those with less rating experience. Third, experienced raters tended to achieve more agreement in their ratings while novice raters tended to have more variations in their assessment perceptions and their application of the rating scale. One possibility to explain this rating behaviour is that the experienced raters in the current study were those who attended over 5 training sessions (section 3.5) and rated almost exclusively VSTEP speaking tests. It appeared to be the community of practice that could have facilitated them having a shared understanding of the rating criteria and the norms of the rating community, as suggested by Lamprianou et

al. (2020). The participant raters also reported that the discussions they had with their colleagues in training and in their rating sessions helped them achieve better understanding of the key terms in the rating scale, leading to their increased confidence in using the scale. They also stated that increased experience of rating in the tests helped with increased capacity to pay detailed attention to the performances during their rating process. The interaction within a community of practice seems to explain why experienced raters have more consistent rating behaviour than novice raters. In addition, the current study provides no evidence that experienced raters can score faster than those with less rating experience as suggested by Sakyi (2003). The confidence of the experienced raters in the study seemed not to correlate with the speed at which they arrived at their scores, as they were all found to award scores at the end of the performance. This finding supports what Davis (2012) revealed in his study that more proficient raters took longer to make decisions. One possibility for this rating behaviour might be that raters with more rating experience were more careful in the way they made their scoring decisions. It also might be that the rating of communicative speaking performances is highly complicated and requires an extra degree of care, rather than being automatised.

Moreover, there was evidence in my study showing that novice raters referred to the rating scales and relied on the criteria more frequently than those with more experience, as suggested by Esfandiari and Noor (2018). However, all of the participants appeared to rely on the criteria listed in the rating scale while rating, which contrasts with an earlier finding (Shirazi, 2012) that "when raters talk, rubric falls silent". One explanation for this finding might lie in the particular context in which the participants were working and the perceptions the participants had towards their rating job. The participant raters were aware of the potential impact of the scores on the TTs, as Rose said: "*it [the test's score] affects their [TTs'] future and it's the final exam, the standard exit exam which*

decides their whole future". The understanding of the significance of the scores and the nurturing aspect of a teacher may have directly affected the way the raters rated. In this case, the raters relied on the rating scale they used and through which they could achieve "the same mindset" (Iris). Thus, one's rating behaviour should not be interpreted as being indicative of a similarity in every rater. Rater characteristics need to be specified and included in studies trying to understand rating behaviour as suggested in H. J. Kim's (2015) study.

7.2.2 What are the factors that cause disagreement among all the raters in making score decisions?

One of the main findings of this research question was that the raters' attitudes toward 'nativeness' and local language features appeared to influence their score decisions. The concept of 'intelligibility' seemed to be perceived with reference to 'nativeness' by the participants and consequently operationalised in such a way. The participant raters were more lenient in awarding pronunciation scores when rating English L1 like speech and more severe when rating Vietnamese accented speech. This finding seems to provide empirical evidence to support the speculation made by O. Kang et al. (2019) about the impact of listeners' attitude towards speakers' pronunciation, an aspect about which Harding (2017) called for further investigation, in order to be widely understood. Prior studies (Zhang & Elder, 2011) have noted that Chinese-L1 raters may be more oriented to standard English than English L1 raters in their ratings. The data in the current study not only further supports this idea but also unpacks its underlying reason. The reason why the participant raters appeared to perceive 'intelligibility' and 'natural pronunciation' the way they did (see section 5.2) was due to the context they were in. In other words, the English standards which they often used in their learning and teaching materials during the past 20 years were either British or American. I would speculate that the

inner-circle English pronunciation is their only choice of standards in Vietnam's context. This aspect of the impact of the broader context of how EFL raters construct their perception of pronunciation in general and form their attitudes towards English standards has been an under-researched area in the field of language testing (see section 2.6). Thus, this study is among the first studies which have provided insights into the issue of how English pronunciation assessment practice is conducted in an EFL context, since other studies (Hsu, 2016, 2019) investigated raters' attitudes in English speaking contexts. It emphasises the importance of understanding the context and the perceptions of the people in that context before any assessment related decisions are made in terms of a glocalised test of English; for example, to define the "intelligibility construct". A considerable degree of caution should be therefore exercised when interpreting research findings about intelligibility and applying them to the field of assessment. What sounds 'intelligible' and 'natural' to raters may not be intelligible and natural to local language users and may not always equate with communicative effectiveness.

On the question of the factors causing disagreement among all the raters, the treatment of local language features in a glocalised test of English was found to be influential. The current study found that VSTEP raters seemed to be more severe when rating Vietnamese accented speech and they did not have positive attitudes towards some local language features identified in the TTs' speech such as the lack of contrast between /l/ and /n/ or the lack of ending sounds. This finding was also reported by Caban (2003) (see section 2.2). This finding is significant as it provides insight into what may contribute to score variations in the Vietnamese context and explains the reason behind raters' severity/leniency, an aspect that quantitative oriented studies fail to unpack. In a similar context, traces of several Japanese language features are treated as acceptable in EIKEN tests in the rating process. It means EIKEN raters can be

lenient in deciding the pronunciation scores of Japanese accented speech as long as they can understand the speech. This sensible decision was made due to the fact that the test is used in the context where Japanese learners mainly use English to interact with Japanese learners in a Japanese academic context. It is interesting to note that EIKEN tests are rated by English L1 speakers with EFL teaching experience, whereas VSTEP tests are rated by Vietnamese EFL teachers. It is, however, important that further research be conducted to verify, refine, and localise the criteria to ensure that the decision has enabled the positive intended impact in its local context. Elder and Harding (2008) reminded us to consider the views of TTs themselves. They observed that although the arguments for contextually sensitive tests are persuasive, "test users in those contexts are often the first to reject them for a range of reasons," including the lack of mobility potential (Blommaert, 2010).

The next finding, which I considered significant, was the raters' orientation toward syntactic and lexical accuracy and complexity. Very little was found in the literature on the question of how human raters rate accuracy and the complexity of aspects of grammar and vocabulary and how much of accuracy and complexity raters take into account when deciding on a speaking performance at a certain level, e.g.: B1, B2 or C1 (see section 2.7). A number of indices are proposed to measure these aspects; however, the indices are mainly based on quantitative measurement. Given the time limit and a number of other cognitive tasks a human rater has to deal with, whether a human rater can utilise these indices in his/her ratings remains unclear. In terms of syntax, this study showed that more severe raters tended to punish mistakes; more lenient raters tended to reward attempts to use complex grammatical structures. The study provided evidence that the raters seemed to use subordination-based variables and/or coordination-based variables, two among four variables suggested by Norris and Ortega (2009) in assessing grammar complexity.

Although several raters (e.g.: Tulip, Sunflower) quantified the number of mistakes and complex structures, it was unclear how they decided “a range” or “a wide range”. This result may be explained by the fact that the language used in the rating scale allowed flexibility to interpret and the rating guidelines may not have specified the quantification required for each level of proficiency. Moreover, in general, the raters appeared to pay more attention to grammatical accuracy. This finding is interesting in itself in the current context where the use of English in the outer circles (Kachru, 1992) is assumed to serve the purpose of getting the message across (Jenkins, 2000). This further supports the necessity of understanding what people in the outer circles perceive and do in relation to the use of standard English and English varieties.

In terms of lexical resource, one important subtheme which emerged from the data was the way the raters assessed lexical sophistication, one of four aspects of assessing vocabulary as suggested by Read (2000) (see section 2.7). The raters seemed to base their ratings on lexical frequency profiles, the academic word list and the English Profile project – vocabulary to pick up “unusual or advanced words”. Raters with more rating experience and teaching experience were more confident in their rating of this aspect. There was no evidence showing that the raters assessed lexical sophistication using all the dimensions suggested by Eguchi and Kyle (2020). This result is important because it unpacks the way human raters assess vocabulary, which is not similar to what has been suggested in the literature. It could be argued that this result was due to the content of the training the raters received; thus, it needs to be interpreted with caution. Furthermore, in the current study, the raters did not have difficulty in identifying good words/phrases, but the problem appeared to lie in the question of how many were sufficient for one level of proficiency. The data indicated that the raters had different ways of solving this issue, which led to score variations. It could be the case that research into vocabulary assessment has been one step

further forward than research into rater cognition. This finding seems to echo the issue mentioned in the literature review that more research is needed to investigate vocabulary measures across levels by human raters.

Regarding discourse management, the considerable amount of attention paid to this assessment criterion seemed not to be due to the difference in the raters' rating experience, but to individual perceptions of assessing the criterion. It could be because this required the raters to pay more attention to making sense of the speech and evaluating its coherence by attending to how the topics were developed, whether the ideas were relevant and how the ideas were linked together. The constructs assessed here may be more complicated than those in other rating criteria that only required the raters to list the errors in grammar and vocabulary and the good words or good complex structures that the TTs could use. The rating experience in itself, thus, may not play an important role in assessing this criterion. Perhaps, it could have been their experience of learning the language that played a role in this case. Being learners, the raters might have understood the significance of getting the message across with a high degree of clarity and the way to achieve this clarity.

Drawing on Knoch et al.'s (2021) model of factors influencing rating quality in rater-mediated assessment, the complex interaction between the variables suggested in the model (rater background, rating experience, rater cognition, community of practice and raters' perceptions of the rating criteria) was seen in the current study. For example, raters with more teaching experience of low-level proficiency learners tended to identify grammatical mistakes more easily than those without such experience. Among the factors suggested, rater training, rating experience and CoP tended to be the most influential for the participant raters. However, one more factor that could be added to the model was the social context the raters were in – in this case this was an EFL context

where teaching and learning materials contributed to shaping the raters' perspectives of English pronunciation in assessment. In this regard, phenomenology was helpful in unpacking this factor by looking closely at the lived experience of the raters. Moreover, the study provides rich, contextual insights into how the raters perceived and applied the rating criteria, an area which is not discussed in detail in Knoch et al. (2021).

7.2.3 In what ways do raters develop their rating practice?

This qualitative study found that the participant raters, regardless of teaching experience, experienced three main stages in developing their rating practice. This study is among the first which reveals the stages of how novice raters evolve into experienced ones. As argued in the literature review (see section 2.4), it is important to unpack how scoring behaviours might evolve and what scoring behaviours are related to experienced raters and what scoring behaviours are associated with novice and developing raters. The results in the current study have provided empirical evidence to extend understanding of this issue.

One of the most significant findings was that the understanding of the script and the rating process gained from the training did not prepare the raters sufficiently to perform all of the required tasks smoothly when they started their rating job. This result is in agreement with Lim's (2011) finding which showed that some novice raters may not be able to show rating consistency after a few months of rating. This again highlights the need to reconsider the follow-up support to novice raters who are newly certified.

Another important finding was that the community of practice, including the amount of training and the volume of ratings completed, played a significant part in developing rating practice. This study supports evidence from previous

investigations (e.g.: Davis, 2016; Lim, 2011). It confirms that rating practice does not evolve with practice (Sahan & Razi, 2020), but instead, the more training programmes the raters attend, the more confident they become in their ratings. This study also confirms that training and interaction with other members in the community helps not only new members develop understanding of the norms and standards the community holds, but also helps experienced raters refine their own learning (Herbert et al., 2014).

7.3 Implications

The study has several implications. First, it is clear that rater characteristics need to be included in studies trying to understand rating behaviour as suggested in H. J. Kim's (2015) study. Specifically, the unequal attention paid to certain assessment criteria might have resulted from the particular teaching experiences of particular raters and the context in which they worked, even though they were all trained and certified. Thus, this rating behaviour should not be interpreted as a similarity of every rater. Second, the findings of this study provide more insights into the strategies used by the raters while they were making their scoring decisions. Holistic rating seems not to be mentioned in the rating with an analytic rating scale; thus, it is important to consider first whether referencing to holistic rating is a construct relevant factor and then whether this is a more or less desirable rating behaviour.

The results of this study also enhance our understanding of local raters' perceptions of the local test-takers' spoken performances in their local context. The results of the research project have several implications for the localised speaking test constructs assessed in an EFL context. The results revealed the significance of identifying the features of the test takers in the local area and the importance of discussing and agreeing on how those features should be treated in the test. Moreover, in an EFL context, the discussion of which standard is

desirable and how it is expected to be treated in the test is vitally important. Those decisions require attention not only from the test developers but also from other stakeholders who accept the test results and education policy makers, as suggested by J. D. Brown (2014), to achieve a balance between locality and globality.

The results of this study also have several implications for enhanced rater training programmes. First, as the study documented the mental processes the raters used in assigning scores, it seems that raters with different backgrounds displayed different needs in their rating processes. Thus, it might be helpful to provide them with individualised feedback to help them become more confident in performing their rating job, rather than one-size-fits-all training programmes for all raters. Second, an agreement as to how different types of lexical and syntactic mistakes should be treated in the context of speaking assessment, compared with a writing assessment context, should be reached before live ratings. This is particularly important for those raters for whom English is not their first language. Third, a clearer explanation, together with point-by-point demonstration of multilayered descriptors using adjectives/adverbs, may be helpful for raters. Additionally, more training on CEFR levels may be needed, with the focus on B2 and C1 features and the agreed quantification of how many complex syntactic and lexical items are sufficient for each level. These can help raters enhance the accuracy of their own sense of the test takers' overall proficiency level. Furthermore, as the study revealed that beginning raters might not have been ready to perform their ratings, it is important to provide follow-up support to newly certified raters during their transition from training to practice.

The research project also has several implications to enhance clarity in the rating scale. First, the issues of how "natural pronunciation" should be conceptualised and operationalised in the rating scale are of significance. The data revealed that

this term allowed flexibility of understanding and subsequently different application among the participant raters in assessing speaking performances. Second, it was observed from the data that the participant raters assessed spoken grammar based on the rules of written grammar. It is important that there is clarification for the raters about which approach is expected to reflect the constructs assessed in VSTEP speaking tests.

This study also contributes to a better understanding of the extent to which English(es) was assessed in the eyes of local raters in their local practice. These insights are hopefully helpful to other local and international testing bodies who are delivering or will be delivering localized standardized tests to local communities elsewhere.

7.4 Limitations

This study was significant as it provided data about the rating process experienced by VSTEP speaking raters and the way they developed their rating practice. This qualitative study was informed by data triangulation and a detailed, rich description of the research process, my insider perspective, and reflexivity to validate the data analysis (see sections 3.9 and 3.10). However, the study has a number of limitations that need to be acknowledged. First, the data set gained from the moderation discussion would have provided more insights if all of the raters were able to attend, and I was aware of my potential influence when I conducted the moderation discussion with several raters (see sections 3.7 and 3.10). However, this study can be considered as an example of how moderation discussions may provide insight into perceptions, practices and experiences regarding the research phenomenon.

Second, the translation of data from Vietnamese to English can be considered as another limitation to the process, as it is not without drawbacks. However,

thanks to the Assessment Research Award achieved from the British Council, a considerable amount of work was done under the strict scrutiny of a translation agency. I am confident that the translated data faithfully represents the meanings expressed by the participants.

7.5 Suggestions for further research

The findings and implications of the study suggest a number of possible avenues for future research. First, since VSTEP is the first standardised test of English proficiency in Vietnam, a replication of this study in the south of Vietnam would be of great value to gain more insight and understanding of the factors which can impact raters' scoring decisions and how they develop their rating practice. The social context of the south of Vietnam might be different from the north where the current study was conducted. Moreover, phenomenology, the chosen qualitative strategy and the use of the three research instruments (observation of moderation discussion, TAPs and interview) was effective in providing significant insights to further understanding of the research phenomenon. Second, it is also important to conduct further research with the aim of extending understanding of the TTs and other stakeholders regarding their perceptions of standard English and English varieties and what VSTEP speaking scores mean to them. These insights might inform the adjustment or development of the test construct and elicit positive intended consequences. Third, future researchers could observe changes in VSTEP rater training proposed by this study and conduct further studies to investigate if these changes would be beneficial to raters. Finally, beyond Vietnam's borders, replication of this study elsewhere could further develop understanding of the mental rating process experienced by speaking raters, the strategies they use to arrive at their final scores and the stages to describe how they develop their

rating practice. This could develop the picture of speaking rater cognition more comprehensively.

7.6 Contributions to knowledge

The contribution of the current study to research and practice derives in part from the substance of its findings, and in part from its methodology. I suggest three areas of potential import:

1) The study provides detailed insights into the mental process of raters rating speaking performances. As a phenomenological study, it offers rich contextual detail, enabling the rating processes of 14 raters and how their rating practice developed to be understood in its full complexity and singularity. The level of detail of the rating and development processes analysed in the study may allow those working in similar contexts to recognise aspects of their own experience in the accounts and relate their experience to the findings of the study.
2) The study offers extended understanding of the factors influencing the raters' decisions of scores. Apart from the factors suggested in Knoch et al.'s (2021) model (section 2.3), the social context in which the raters live should be considered to achieve better understanding of the issue of why raters behave the way they do. The study reveals how the broader context contributed to shaping the way the raters perceived and conducted their assessment.
3) The methodological contributions lie in the study's use of phenomenology and IPA to explore the construction of the rating processes and rating development processes. The use of IPA in the study allowed the processes to surface. The detailed analysis of each case and the recurrent themes demonstrated the ability to illuminate valuable information about rating practice and its context. The current study on exploring the lived experience

of raters performing their rating job has shown that observing the raters while they were doing their job (moderation discussion), listening to the raters' inner thinking (TAPs) and asking them about their experience can enlighten our understanding of the rating processes, of the factors that impact their score decisions and of their rating development process

7.7 Conclusion

The aims of the current study were to illuminate the lived experience of the raters within the context of a speaking performance test by looking closely at their mental rating processes, unpacking the factors influencing their score decisions and revealing the stages of their rating practice development. Conducting research which starts from the experiences of those who work within it is particularly valuable. If we are to support our raters in this era of high stakes, we should acknowledge the voices of our raters who are scoring in high stakes tests. It is hoped that the results of this study will contribute to a better understanding of what it means to be a rater rating a speaking performance in a high-stakes test.

References

[BERA], B. E. R. A. (2018). Ethical guidelines for educational research. In. London.

[ILTA], I. L. T. A. (2018). Codes of ethics. In ILTA (Ed.). Vancouver.

Al-Maamari, F. (2016). Community of assessment practice or interests: The case of EAP writing assessment. *Indonesian journal of applied linguistics, 5*(2), 272-281. doi:10.17509/ijal.v5i2.1351

Aldridge, A., & Levine, K. (2001). *Surveying the social world: principles and practice in survey research*. Buckingham;Philadelphia, PA;: Open University Press.

Anh, D. (2018, 09 November). Hà Nội xóa nói ngọng trong nhà trường: Kết quả không thể đến trong ngày một ngày hai. *An ninh thủ đô*. Retrieved from https://anninhthudo.vn/ha-noi-xoa-noi-ngong-trong-nha-truong-ket-qua-khong-the-den-trong-ngay-mot-ngay-hai-post374060.antd

Babbie, E. R. (2015). *The practice of social research*: Nelson Education.

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

Bachman, L., & Palmer, A. S. (1981). The construct validation of the FSI oral interview *Language Learning, 31*(1), 67-86.

Bachman, L., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*: Oxford University Press.

Bachman, L. F. (2004). *Statistical Analyses for Language Assessment Book*: Cambridge University Press.

Baird, J.-A., Greatorex, J., & Bell, J. F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in education : principles, policy & practice, 11*(3), 331-348. doi:10.1080/0969594042000304627

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

Baker, E. (1995). *Validity and equity issues in educational assessment.* Paper presented at the 17th annual Language Testing Research Colloquium, Long Beach, CA.

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

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

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. doi:10.1080/0969594X.2010.526585

Barrett, S. (2001). The impact of training on rater variability. *International Education Journal, 2*(1), 49-58.

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

Bell, L. (2014). Ethics and feminist research. In *Feminist research practice: A primer* (pp. 73-106).

Blommaert, J. (2010). *The sociolinguistics of globalization*: Cambridge University Press.

Bodewig, C., Badiani-Magnusson, R., Macdonald, K., Newhouse, D., & Rutkowski, J. (2014). *Skilling up Vietnam: Preparing the workforce for a modern market economy*: World Bank Publications.

Boers, F., Demecheleer, M., & Eyckmans, J. (2004). Cross-cultural Variation as a Variable in Comprehending and Remembering Figurative Idioms. *European journal of English studies, 8*(3), 375-388. doi:10.1080/1382557042000277449

Bøhn, H., & Hansen, T. (2017). Assessing Pronunciation in an EFL Context: Teachers' Orientations towards Nativeness and Intelligibility. *Language Assessment Quarterly, 14*(1), 54-68. doi:10.1080/15434303.2016.1256407

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. doi:10.1111/lang.12067

Briggs, C. L. (1986). *Learning how to ask: a sociolinguistic appraisal of the role of the interview in social science research*. Cambridge: Cambridge University Press.

Brinkmann, S. (2013). *Qualitative interviewing*. Oxford, England: Oxford University Press.

Brown, A. (2000). *An investigation of the rating process in the IELTS Oral interview*. Retrieved from

Brown, A. (2006). *An examination of the rating process in the revised IELTS Speaking Test*. Retrieved from

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test taker performance on English for Academic Purposes speaking tasks*. Retrieved from Educational Testing Service:

Brown, A., & McNamara, T. (2004). "The Devil Is in the Detail": Researching Gender Issues in Language Assessment. *TESOL Quarterly, 38*(3), 524-538. doi:10.2307/3588353

Brown, A., & Taylor, L. (2006). *A worldwide survey of examiners' views and experience of the revised IETLS Speaking Test*. Retrieved from

Brown, J. D. (2014). The Future of World Englishes in Language Testing. *Language Assessment Quarterly, 11*(1), 5-26. doi:10.1080/15434303.2013.869817

Bryman, A. (2008). Of methods and methodology. *Qualitative Research in Organizations and Management: An International Journal, 3*(2), 159-168. doi:10.1108/17465640810900568

Buck, S., Ritter, G. W., Jensen, N. C., & Rose, C. P. (2010). Teachers Say the Most Interesting Things - An Alternative View of Testing. *The Phi Delta Kappan, 91*(6), 50-54. doi:10.1177/003172171009100613

Bygate, M. (2001). Effects of task repetition on the structure and control of language. In *Researching Pedagogic Tasks: Second Language Learning, Teaching and Testing*: Longman.

Caban, H. L. (2003). Rater group bias in the speaking assessment of four L1 Japanese ESL students. *University of Hawai'I Second Langauge Studies Paper 21 (2)*.

Cameron, C. A., Lee, K., Webster, S., Munro, K., Hunt, A. K., & Linton, M. J. (1995). Text cohesion in children's narrative writing. *Applied Psycholinguistics, 16*(3), 257-269.

Celce-Murcia, M., Dörnyei, Z., & Thurrell, S. (1995). Communicative competence: A pedagogically motivated model with content specifications. *Issues in Applied linguistics, 6*(2), 5-35.

Cheng, L. (2008). The key to success: English language testing in China. *Language Testing, 25*(1), 15-37.

Clark, H. H., & Tree, J. E. F. (2002). Using uh and um in spontaneous speaking. *Cognition, 84*(1), 73-111.

Coffman, W. E. (1971). On the Reliability of Ratings of Essay Examinations in English. *Research in the teaching of English, 5*(1), 24-36.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (6th ed.). London: Routledge.

Cohen, L., Manion, L., & Morrison, K. (2011). *Research methods in education* (7th;Seventh; ed.). London: Routledge.
Cohen, L., Manion, L., Morrison, K., & Bell, R. (2011). *Research methods in education* (7th ed.). London: Routledge.
Connor-Linton, J. (1995). Looking Behind the Curtain: What Do L2 Composition Ratings Really Mean? *TESOL Quarterly, 29*(4), 762-765. doi:10.2307/3588174
Costigan, A. T. (2002). Teaching the Culture of High Stakes Testing: Listening to New Teachers. *Action in Teacher Education, 23*(4), 28-34. doi:10.1080/01626620.2002.10463085
Creswell, J. W. (2013). *Qualitative inquiry & research design: choosing among five approaches* (Third ed.). London;Los Angeles, [Calif.];: SAGE.
Creswell, J. W. (2014). *Research design: qualitative, quantitative, and mixed methods approaches* (Fourth, International student ed.). Los Angeles: SAGE.
Creswell, J. W., & Creswell, J. D. (2018). *Research design: qualitative, quantitative & mixed methods approaches* (5th, international student ed.). Los Angeles: SAGE.
Creswell, J. W., & Miller, D. L. (2000). Determining validity in qualitative inquiry *Theory Into Practice, 39*(3), 124-130.
Crisp, V. (2012). An Investigation of Rater Cognition in the Assessment of Projects. *Educational measurement, issues and practice, 31*(3), 10-20. doi:10.1111/j.1745-3992.2012.00239.x
Crossley, S. A., Salsbury, T., & Mcnamara, D. S. (2015). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics, 36*(5), 570-590.
Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing, 28*(4), 561-580.
Crotty, M. (1998). *The foundations of social research: meaning and perspective in the research process*. Los Angeles: Saga.
Crusan, D. (2015). Dance, ten; looks, three: Why rubrics matter. *Assessing Writing, 26*, 1-4. doi:10.1016/j.asw.2015.08.002
Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*(1), 31-51.
Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision Making While Rating ESL/EFL Writing Tasks: A Descriptive Framework. *The Modern Language Journal, 86*(1), 67-96. doi:10.1111/1540-4781.00137
Cushing, S. (2019). [Speaking and Writing rater training].
Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing* (Vol. 7): Cambridge University Press.
Davis, L. (2012). *Rater expertise in a second language speaking assessment: The influence of training and experience.* University of Hawai'i at Manoa,
Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117-135. doi:10.1177/0265532215582282
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. doi:10.1191/0265532204lt286oa
de Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching, 54*(2), 113-132.
de Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). FACETS OF SPEAKING PROFICIENCY. *Studies in Second Language Acquisition, 34*(1), 5-34. doi:10.1017/S0272263111000489

Denscombe, M. (2014). *Good Research Guide: For Small-Scale Social Research Projects* (5th;Fifth; ed.). :: McGraw-Hill Education.
Denzin, N. K., & Lincoln, Y. S. (2000). *The handbook of qualitative research* (2nd ed.). London: SAGE.
DeRemer, M. L. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing, 5*(1), 7-29.
Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second Language Fluency: Judgments on Different Tasks: Language Learning. *Language Learning, 54*(4), 655-679. doi:10.1111/j.1467-9922.2004.00282.x
Deterding, D. (2010). Norms for pronunciation in Southeast Asia. *World Englishes, 29*(3), 364-377. doi:10.1111/j.1467-971X.2010.01660.x
Dimova, S. (2017). Pronunciation Assessment in the Context of World Englishes. In O. Kang & A. Ginther (Eds.), *Assessment in second language pronunciation*: Taylor & Francis.
Dunlea, J. (2013). *Recognition for locally developed tests: An overview of regional approaches*. Paper presented at the The 1st Brtish Council New Directions in English Language Assessment conference, Beijing, China.
Dunlea, J., Nguyen, T. N. Q., Spiby, R., Nguyen, T. Q. Y., Nguyen, T. M. H., Nguyen, T. P. T., & Thai, H. L. T. (2017). *A multi-method study to investigate the constructs measured by two EFL tests and their comparability in the context of Vietnam*. Paper presented at the The 4th International Conference of the Asian Association for Language Assessment, Taipei, Taiwan.
Dunlea, J., Spiby, R., Nguyen, T. N. Q., Nguyen, T. Q. Y., Nguyen, T. M. H., Nguyen, T. P. T., . . . Bui, T. S. (2018). *APTIS-VSTEP Comparability Study: Investigating the usage of two EFL Tests in the context of Higher Education in Vietnam*. Retrieved from British Council:
Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155-185.
Eckes, T. (2012). Operational Rater Types in Writing Assessment: Linking Rater Cognition to Rater Behavior. *Language Assessment Quarterly, 9*(3), 270-292. doi:10.1080/15434303.2011.649381
Eguchi, M., & Kyle, K. (2020). Continuing to Explore the Multidimensional Nature of Lexical Sophistication: The Case of Oral Proficiency Interviews. *The Modern language journal (Boulder, Colo.), 104*(2), 381-400. doi:10.1111/modl.12637
Elder, C., & Davies, A. (2006). ASSESSING ENGLISH AS A LINGUA FRANCA. *Annual Review of Applied Linguistics, 26*, 282-304. doi:10.1017/S0267190506000146
Elder, C., & Harding, L. (2008). Language testing and English as an international language: Constraints and contributions. *Australian Review of Applied Linguistics, 31*(3), 34.31-34.31.
Ellis, R., & Barkhuizen, G. (2005). *Analysing Learner Language*: OUP Oxford.
Emerson, R. M., Fretz, R. I., & Shaw, L. L. (2001). Participant observation and fieldnotes. In *Handbook of ethnography* (pp. 352-368).
Erickson, F. (2018). A History of Qualitative Inquiry in Social and Educational Research. In N. K. Denzin & Y. S. Lincoln (Eds.), *The SAGE handbook of qualitative research* (Fifth ed.). Thousand Oaks;Los Angeles;: Sage Publications, Ltd.
Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: verbal reports as data* (Rev. ed.). London;Cambridge, Mass;: 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. doi:10.30479/jmrels.2019.9383.1248

Europe, C. f. C. C.-o. E. C. C. o. (2001). *Common European framework of reference for languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.
Field, J. (2005). Intelligibility and the Listener: The Role of Lexical Stress. *TESOL Quarterly, 39*(3), 399-424.
Fillmore, L. W. (1979). Individual differences in second language acquisition. In *Individual differences in language ability and language behavior* (pp. 203-228): Elsevier.
Finlay, L. (2008). A Dance Between the Reduction and Reflexivity: Explicating the "Phenomenological Psychological Attitude". *Journal of Phenomenological Psychology, 39*(1), 1-32. doi:10.1163/156916208X311601
Finlay, L. (2009). Debating Phenomenological Research Methods. *Phenomenology & Practice, 3*(1). doi:10.29173/pandpr19818
Finlay, L., & Gough, B. (2003). *Reflexivity: a practical guide for researchers in health and social sciences* (1. Aufl.;1st; ed.). Malden, MA: Blackwell Science.
Foster, P., & Skehan, P. (1996). The Influence of Planning and Task Type on Second Language Performance. *Studies in Second Language Acquisition, 18*(3), 299-323. doi:10.1017/S0272263100015047
Fulcher, G. (1993). *The construction and validation of rating scales for oral tests in English as a foreign language.* Citeseer,
Fulcher, G. (2003). *Testing second language speaking*: Pearson Education.
Ga, V. B. (2017). *Regulation of organising Vietnamese standardised test of English proficiency*. Ministry of Education and Training, Vietnam
Gass, S. M. (2001). Innovations in second language research methods. *Annual Review of Applied Linguistics, 21*(1), 221-232. doi:10.1017/S0267190501000137
Geertz, C., & Societies, A. C. o. L. (1973). *The Interpretation Of Cultures*: Basic Books.
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.
Giorgi, A., & Giorgi, B. (2008). Phenomenology. In J. A. Smith (Ed.), *Qualitative psychology: a practical guide to research methods* (2nd ed.). London;Los Angeles, Calif;: SAGE.
Glesne, C. (2016). *Becoming qualitative researchers: an introduction* (Fifth ed.). Boston: Pearson.
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.
Goldberg, G. L. (2012). Judgment-Based Scoring by Teachers as Professional Development: Distinguishing Promises from Proof. *Educational Measurement: Issues and Practice, 31*(3), 38-47. doi:10.1111/j.1745-3992.2012.00242.x
Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers, 36*(2), 193-202.
Green, C., & Gilhooly, K. (1996). Protocol analysis: practical implementation. In J. T. Richardson (Ed.), *Handbook of qualitative research methods for psychology and the social sciences* (pp. 55 - 74): BPS books.
Groves, J. M., & Chan, H. T. (2010). Lexical traps in Hong Kong English. *English Today, 26*(4), 44-50. doi:10.1017/S0266078410000337
Guba, E. G. (1990, 1990). *The Paradigm dialog*, Newbury Park, Calif;London;.
Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. In *Handbook of qualitative research* (Vol. 2, pp. 105).
Guillot, M.-N. (1999). *Fluency and its teaching* (Vol. 11): Multilingual Matters.

Halliday, M. A. K. (1985). *An introduction to functional grammar*. London: Edward Arnold.
Halliday, M. A. K., & Matthiessen, C. M. I. M. (2004). *An introduction to functional grammar* (3rd ed.). London: Arnold.
Hammersley, M. (2008). *Questioning qualitative inquiry: critical essays*. London: SAGE.
Hammersley, M. (2013). *What is qualitative research?* (1 ed.). London;New York;: Bloomsbury.
Hammersley, M., & Atkinson, P. (2007). *Ethnography: principles in practice* (3rd ed.). London;New York;: Routledge.
Han, Q. (2016). *Rater Cognition in L2 Speaking Assessment : A Review of the Literature*.
Harding, L. (2012). Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective. *Language Testing, 29*(2), 163-180. doi:10.1177/0265532211421161
Harding, L. (2016). What Do Raters Need in a Pronunciation Scale? The User's View. In T. Isaacs & P. Trofimovich (Eds.), *Second language pronunciation assessment: Interdisciplinary perspectives*.
Harding, L. (2017). Validity in Pronunciation Assessment. In O. Kang & A. Ginther (Eds.), *Assessment in second language pronunciation*: Taylor & Francis.
Henn, M., Weinstein, M., & Foard, N. (2009). *A critical introduction to social research* (2 ed.). London: SAGE.
Hennink, M. M., Hutter, I., & Bailey, A. (2011). *Qualitative research methods*. London;Los Angeles;: SAGE.
Herbert, I. P., Joyce, J., & Hassall, T. (2014). Assessment in Higher Education: The Potential for a Community of Practice to Improve Inter-marker Reliability. *Accounting education (London, England), 23*(6), 542-561. doi:10.1080/09639284.2014.974195
Hill, K. (1996). Who should be the judge? The use of non-native speakers as raters on a test of English as an international language. *Melbourne Papers in Language Testing, 5*(2), 29-50.
Hsieh, C.-N. (2011). *Rater Effects in ITA Testing: ESL Teachers' versus American Undergraduates' Judgments of Accentedness, Comprehensibility, and Oral Proficiency*: ERIC.
Hsu, T. H.-L. (2016). Removing bias towards World Englishes: The development of a Rater Attitude Instrument using Indian English as a stimulus. *Language Testing, 33*(3), 367-389. doi:10.1177/0265532215590694
Hsu, T. H.-L. (2019). Rater attitude towards emerging varieties of English: a new rater effect? *Language Testing in Asia, 9*(1), 1-21. doi:10.1186/s40468-019-0080-0
Huang, B. H. (2013). The effects of accent familiarity and language teaching experience on raters' judgments of non-native speech. *System, 41*(3), 770-785. doi:10.1016/j.system.2013.07.009
Huang, J. (2008). How accurate are ESL students' holistic writing scores on large-scale assessments?—A generalizability theory approach. *Assessing Writing, 13*(3), 201-218. doi:10.1016/j.asw.2008.10.002
Huhta, A., Alanen, R., Tarnanen, M., Martin, M., & Hirvelä, T. (2014). Assessing learners' writing skills in a SLA study: Validating the rating process across tasks, scales and languages. *Language Testing, 31*(3), 307-328.
Inoue, C. (2016). A comparative study of the variables used to measure syntactic complexity and accuracy in task-based research. *Language learning journal, 44*(4), 487-505. doi:10.1080/09571736.2015.1130079
Inoue, C., Khabbazbashi, N., Lam, D. M., & Nakatsuhara, F. (2021). *Towards new avenues for the IELTS Speaking Test: insights from examiners' voices* (2201-2982). Retrieved from
Isaacs, T. (2014). Assessing Pronunciation. In A. J. Kunnan (Ed.), *The Companion to Language Assessment* (pp. 140-155). Hoboken, NJ: Wiley-Blackwell.

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. doi:10.1080/15434303.2013.769545

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*(1), 24-49.

Iwashita, N., May, L., & Moore, P. (2017). Features of discourse and lexical richness at different performance levels in the APTIS speaking test (AR-G/2017/2).

Iwashita, N., McNamara, T., & Elder, C. (2001). Can We Predict Task Difficulty in an Oral Proficiency Test? Exploring the Potential of an Information-Processing Approach to Task Design. *Language Learning, 51*(3), 401-436. doi:10.1111/0023-8333.00160

Iwashita, N., & Vasquez, C. (2015). *An examination of discourse competence at different proficiency levels in IELTS Speaking Part 2*. Retrieved from

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning, 63*, 87-106.

Jenkins, J. (2000). *The phonology of English as an international language*: Oxford university press.

Jenkins, J. (2006). Current Perspectives on Teaching World Englishes and English as a Lingua Franca. *TESOL Quarterly, 40*(1), 157-181. doi:10.2307/40264515

Jones, R. L. (1979). The oral interview of the foreign service institute. In B. Spolsky (Ed.), *Some major tests*. Wasington, DC: Center for Applied Linguistics.

Kachru, B. B. (1992). World Englishes: Approaches, issues and resources. *Language Teaching, 25*(1), 1-14.

Kang, J. Y. (2005). Written narratives as an index of L2 competence in Korean EFL learners. *Journal of second language writing, 14*(4), 259-279. doi:10.1016/j.jslw.2005.10.002

Kang, O., Rubin, D., & Kermad, A. (2019). The effect of training and rater differences on oral proficiency assessment. *Language Testing, 36*(4), 481-504. doi:10.1177/0265532219849522

Keen, E. (1982). *A primer in phenomenological psychology*. Washington, D.C: University Press of America.

Kenyon, D. (1992). *Introductory remarks at symposium on development and use of rating scales in language testing.* Paper presented at the 14th language testing research colloquium, Vancouver.

Kim, H. J. (2015). A Qualitative Analysis of Rater Behavior on an L2 Speaking Assessment. *Language Assessment Quarterly, 12*(3), 239-239. doi: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. doi:10.1177/0265532208101010

King, N. (2010). Research ethics in qualitative research. In M. A. Forrester (Ed.), *Doing qualitative research in psychology: a practical guide* (pp. 98-118). London;Los Angeles, [Calif.];: SAGE.

Klenowski, V., & Wyatt-Smith, C. (2012). The impact of high stakes testing: the Australian story. *Assessment in Education: Principles, Policy & Practice, 19*(1), 65-79. doi:10.1080/0969594X.2011.592972

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

Knoch, U. (2010). Investigating the effectiveness of individualized feedback to rating behaviour - a longitudinal study. *Language Testing, 28*(2), 179-200.

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior — a longitudinal study. *Language Testing, 28*(2), 179-200. doi:10.1177/0265532210384252

Knoch, U., Fairbairn, J., & Jin, Y. (2021). *Scoring second language spoken and written performance: issues options and directions*: Equinox Publishing Ltd.
Koizumi, R., & In'nami, Y. (2013). Vocabulary knowledge and speaking proficiency among second language learners from novice to intermediate levels. *Journal of Language Teaching & Research, 4*(5).
Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32*(2), 145-164. doi:10.1016/j.system.2004.01.001
Kuusela, H., & Paul, P. (2000). A Comparison of Concurrent and Retrospective Verbal Protocol Analysis. *The American Journal of Psychology, 113*(3), 387-404. doi:10.2307/1423365
Kvale, S. (2007). *Doing interviews*. London: SAGE.
Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. *Behavior research methods, 50*(3), 1030-1046. doi:10.3758/s13428-017-0924-4
Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly, 49*(4), 757-786.
Lamprianou, I., Tsagari, D., & Kyriakou, N. (2020). The longitudinal stability of rating characteristics in an EFL examination: Methodological and substantive considerations. *Language Testing*, 26553222094096. doi:10.1177/0265532220940960
Langdridge, D. (2008). Phenomenology and Critical Social Psychology: Directions and Debates in Theory and Research. *Social and Personality Psychology Compass, 2*(3), 1126-1142. doi:10.1111/j.1751-9004.2008.00114.x
Lather, P. (1993). Fertile Obsession: Validity After Poststructuralism. *The Sociological Quarterly, 34*(4), 673-693. doi:10.1111/j.1533-8525.1993.tb00112.x
Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*(3), 307-322.
Lawthom, R., & Tindall, C. (2011). Phenomenology. In P. Banister (Ed.), *Qualitative methods in psychology: a research guide* (2nd ed.). Maidenhead: McGraw-Hill/Open University Press.
Le, C. V. (2019). English language teaching in Vietnam: aspirations, realities and challenges. In C. V. Le, H. T. M. Nguyen, T. T. M. Nguyen, & R. Barnard (Eds.), *Building Teacher Capacity in English Language Teaching in Vietnam: Research, Policy and Practice*: Routledge.
Le, C. V., Nguyen, H. T. M., Nguyen, T. T. M., & Barnard, R. (2019). *Building Teacher Capacity in English Language Teaching in Vietnam: Research, Policy and Practice*: Routledge.
Leavy, P. (2014). *The Oxford handbook of qualitative research*: Oxford library of psychology.
Lennon, P. (1990). Investigating Fluency in EFL: A Quantitative Approach. *Language Learning, 40*(3), 387-417. doi:10.1111/j.1467-1770.1990.tb00669.x
Levis, J. M. (2006). Pronunciation and the Assessment of spoken language. In R. Hughes (Ed.), *Spoken English, Tesol and Applied Linguistics*. Palgrave Macmillan, London.
Li, H., & He, L. (2015). A Comparison of EFL Raters' Essay-Rating Processes Across Two Types of Rating Scales. *Language Assessment Quarterly, 12*(2), 178-178. doi:10.1080/15434303.2015.1011738
Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *28*(4), 543-560. doi:10.1177/0265532211406422
Linacre, J. M. (1988). FACETS: A computer program for the analysis of multi-faceted data. *Computer program Chicago: MESA*.
Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. London;Newbury Park, Calif;: Sage.

Ling, G., Mollaun, P., & Xi, X. (2014). A study on the impact of fatigue on human raters when scoring speaking responses. *Language Testing, 31*(4), 479-499. doi:10.1177/0265532214530699
Lopez, K. A., & Willis, D. G. (2004). Descriptive versus interpretive phenomenology: Their contributions to nursing knowledge. *Qualitative Health Research, 14*(5), 726-735.
Loxley, A., & Seery, A. (2008). Some philosophical and other related issues of insider research. In *Researching education from the inside* (pp. 23-40): Routledge.
Lu, X. (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern language journal (Boulder, Colo.), 96*(2), 190-208. doi:10.1111/j.1540-4781.2011.01232.x
Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing, 19*(3), 246-276.
Lumley, T. (2005). *Assessing Second Language Writing: The Rater's Perspective*: P. Lang.
Lumley, T., & McNamara, T. (1995). Rater characteristic and rater bias: implications for training. *Language Testing, 12*(1), 54-71.
Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.
Lyle, J. (2003). Stimulated recall: a report on its use in naturalistic research. *British Educational Research Journal, 29*(6), 861-878. doi:10.1080/0141192032000137349
Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balasubramanian, C. (2002). The Effects of Nonnative Accents on Listening Comprehension: Implications for ESL Assessment. *TESOL Quarterly, 36*(2), 173-190. doi:10.2307/3588329
Maxwell, J. A. (1996). *Qualitative research design: an interative approach* (Vol. 41). Thousand Oaks, Calif: Sage Publications.
May, L. (2006). An examination of rater orientations on a paired candidate discussion task through stimulated verbal recall. *Melbourne Papers in Language Testing, 11*(1), 29-51.
May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing, 26*(3), 397-421. doi:10.1177/0265532209104668
McNamara, T. (1996). *Measuring Second Language Performance*: Longman.
McNamara, T. (2009). Principles of testing and assessment. In (pp. 607-628).
McNamara, T., & Knoch, U. (2012). The Rasch wars : the emergence of Rasch measurement in language testing. *Language Testing, 29*(4), 555-576. doi:10.1177/0265532211430367
McNamara, T., & Roever, C. (2006). *Language testing: the social dimension*. Malden, MA: Blackwell Pub.
Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition, 20*(1), 83-108. doi:10.1017/S0272263198001041
Merriam, S. B. (2009). *Qualitative research: a guide to design and implementation* (3rd;Rev. and expand; ed.). San Francisco, California: Jossey-Bass.
Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: an expanded sourcebook* (2nd ed.). Thousand Oaks, Calif;London;: Sage.
Moustakas, C. E. (1994). *Phenomenological research methods*. Thousand Oaks, Calif: Sage.
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. doi:10.1093/applin/amr016
Myford, C. M. (2012). Rater cognition research Some possible directions for the future. *Educational Measurement: Issues and Practice, 31*(3), 48-49.
Nakatsuhara, F., Taylor, L., & Jaiyote, S. (2020). The Role of the L1 in Testing L2 English. In *Ontologies of English* (pp. 187-208).

Nation, I. S., & Webb, S. A. (2011). *Researching and analyzing vocabulary*: Heinle, Cengage Learning Boston, MA.

Ngo, X. M. (2018). Sociopolitical contexts of EFL writing assessment in Vietnam: Impact of a national project. In *The politics of English second language writing assessment in global contexts* (pp. 47-59): Routledge.

Nguyen, H. (2014). *Technical report on the development of test format and specification of Vietnamese Standardised Test of English Proficiency levels 3 to 5 for post-secondary English learners*. Retrieved from www.vstep.vn/sites/default/files/17.4.2015_vstep_report.pdf

Nguyen, H. T. M. (2011). Primary English language education policy in Vietnam: Insights from implementation. *Current Issues in Language Planning, 12*(2), 225-249.

Nguyen, H. T. M., Nguyen, H., Do, T. T. H., Tran, T. H. P., Huynh, A. T., Dang, T., & Davidson, F. (2017). *Developing the Vietnamese Standardised Test of English Proficiency*. Paper presented at the The 3rd International Conference on Language Testing and Assessment, Shanghai, China.

Nguyen, N. (2012). How English has displaced Russian and other foreign languages in Vietnam since "Doi Moi". *International Journal of Humanities and Social Science, 2*(23), 259-266.

Nguyen, T. N. Q. (2019). The Introduction of VSTEP in the context of education reform in Vietnam. In L. I. W. Su, C. J. Weir, & J. R. W. Wu (Eds.), *English Language Proficiency Testing in Asia: A New Paradigm Bridging Global and Local Contexts*: Taylor & Francis.

Nguyen, T. N. Q., Nguyen, T. Q. Y., Nguyen, T. P. T., & Carr, T. (2017). *Using multiple approaches to examine the dependability of VSTEP Speaking and Writing assessments*. Paper presented at the The 4th annual international conference of the Asian Association for language assessment, Taipei, Taiwan.

Nguyen, T. N. Q., Nguyen, T. Q. Y., Tran, T. T. H., Nguyen, T. P. T., Bui, T. S., Nguyen, T. C., & Nguyen, Q. H. (2020). The effectiveness OF VSTEP. 3-5 speaking rater training. *VNU Journal of Foreign Studies, 36*(4).

Norris, J. M., & Ortega, L. (2009). Towards an Organic Approach to Investigating CAF in Instructed SLA: The Case of Complexity. *Applied Linguistics, 30*(4), 555-578. doi:10.1093/applin/amp044

O'Loughlin, K. (1995). Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. *Language Testing, 12*(2), 217-237.

O'Sullivan, B. (2012). A brief history of language testing. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge guide to second language assessment*: Cambridge University Press.

O'Sullivan, B., & Rignall, M. (2007). *Assessing the value of bias analysis feedback to raters for the IELTS writing module*. Retrieved from

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

Ortlipp, M. (2008). Keeping and Using Reflective Journals in the Qualitative Research Process. *Qualitative Report, 13*(4), 695.

Paltridge, B. (1994). Genre Analysis and the Identification of Textual Boundaries. *Applied Linguistics, 15*(3), 288-299. doi:10.1093/applin/15.3.288

Patton, M. Q. (2002). *Qualitative research & evaluation methods* (3rd ed.). Thousand Oaks, Calif;London;: Sage.

Patton, M. Q. (2015). *Qualitative research & evaluation methods: integrating theory and practice* (Fourth ed.). London: SAGE.

Phan, A. N. Q. (2021). Under the impacts of globalisation: the rising power of English as a foreign language (EFL) and the corresponding response of EFL policy in Vietnam. *SN Social Sciences, 1*(1), 31. doi:10.1007/s43545-020-00047-9

Pickard, A. J. (2007). *Research methods in information*. London: Facet.

Pinget, A.-F., Bosker, H. R., Quené, H., & de Jong, N. H. (2014). Native speakers' perceptions of fluency and accent in L2 speech. *Language Testing, 31*(3), 349-365. doi:10.1177/0265532214526177

Pollitt, A., & Murray, N. L. (1996). What raters really pay attention to. In N. Saville (Ed.), *Studies in language testing 3: Performance testing, Cognition and Assessment*. Cambridge: Cambridge University Press.

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. doi:10.3138/cmlr.1748

Purpura, J. E. (2013). Cognition and Language Assessment. In A. J. Kunnan (Ed.), *The Companion to Language Assessment*.

Read, J. (2000). *Assessing vocabulary*: Cambridge university press.

Richardson, J. T. (1996). *Handbook of qualitative research methods for psychology and the social sciences*: BPS books.

Ritchie, J., & Lewis, J. (2003). *Qualitative research practice: a guide for social science students and researchers*. London: SAGE.

Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: Effects on L2 speech production, interaction, uptake and perceptions of task difficulty. *International review of applied linguistics in language teaching, IRAL, 45*(3), 193-213. doi:10.1515/iral.2007.009

Robson, C., & McCartan, K. (2016). *Real world research: a resource for users of social research methods in applied settings* (Fourth ed.). Chichester: Wiley.

Rossiter, M. (2009). Perceptions of L2 Fluency by Native and Non-native Speakers of English. *The Canadian Modern Language Review / La revue canadienne des langues vivantes, 65*(3), 395-412. doi:10.3138/cmlr.65.3.395

Rubin, H. J., & Rubin, I. (2012). *Qualitative interviewing: the art of hearing data* (3rd ed.). London;Los Angeles, [Calif.];: SAGE.

Şahan, Ö., & Razı, S. (2020). Do experience and text quality matter for raters' decision-making behaviors? *Language Testing, 37*(3), 026553221990022-026553221990332. doi:10.1177/0265532219900228

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

Saldaña, J. (2011). *Fundamentals of qualitative research*. New York: Oxford University Press.

Sartre, J.-P. (1996). *Being and nothingness: an essay on phenomenological ontology*. London: Routledge.

Schmitt, N. (2009). *Lexical analysis of input prompts and examinee output of Cambridge ESOL Main Suite Speaking tests*. Retrieved from internal report commissioned by Cambridge ESOL

Schwandt, T. A. (1997). *Qualitative inquiry: A dictionary of terms*. Thousand Oaks, CA, US: Sage Publications, Inc.

Seale, C. (1999). Quality in Qualitative Research. *Qualitative Inquiry, 5*(4), 465-478. doi:10.1177/107780049900500402

Seedhouse, P., & Harris, A. (2011). *Topic development in the IELTS Speaking Test*. Retrieved from
Segalowitz, N. (2010). *Cognitive Bases of Second Language Fluency*: Taylor & Francis.
Seidman, I. (2006). *Interviewing as Qualitative Research: A Guide for Researchers in Education and the Social Sciences*: Teachers College Press.
Sewell, A. (2013). Language testing and international intelligibility: A Hong Kong case study. *Language Assessment Quarterly, 10*(4), 423-443. doi:10.1080/15434303.2013.824974
Shaw, R. (2010). QM3: Interpretative Phenomenological Analysis. In M. A. Forrester (Ed.), *Doing qualitative research in psychology: a practical guide* (pp. 177-201). London;Los Angeles, [Calif.];: SAGE.
Shaw, S. (2002). *The effect of training and standardisation on rater judgement and inter-rater reliability*. Retrieved from
Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing* (Vol. 26): Cambridge University Press.
Sheehan, S., & Munro, S. (2017). *Assessment: attitudes, practices and needs*. Retrieved from British Council:
Shenton, A. K. (2004). Strategies for ensuring trustworthiness in qualitative research projects. *Education for information, 22*(2), 63-75.
Shipman, M. D. (1997). *The limitations of social research* (4th ed.). London;New York;: Longman.
Shirazi, M. A. (2012). When raters talk, rubrics fall silent. *Language Testing in Asia, 2*(4), 123-139.
Shohamy, E. G. (2001). *The Power of Tests: A Critical Perspective on the Uses of Language Tests*: Longman.
Simons, H. (2009). Evolution and Concept of Case Study Research. In (pp. 12). London: SAGE Publications, Ltd.
Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics, 17*(1), 38-62.
Skehan, P. (2018). Lexical Performance by Native and Non-Native Speakers on Language-Learning Tasks. In (1 ed., pp. 189-205): Routledge.
Skehan, P., & Foster, P. (1999). The Influence of Task Structure and Processing Conditions on Narrative Retellings. *Language Learning, 49*(1), 93-120. doi:10.1111/1467-9922.00071
Skehan, P., & Foster, P. (2012). Complexity, accuracy, fluency and lexis in task-based performance. In *Dimensions of L2 performance proficiency: Complexity, accuracy, fluency in SLA* (Vol. 32, pp. 199).
Skehan, P., & Shum, S. (2017). What Influences Performance?: Personal style or the task being done? In *Faces of English Education* (pp. 28-43): Routledge.
Smith, J. A. (2008). *Qualitative psychology: a practical guide to research methods* (2nd ed.). London;Los Angeles, Calif;: SAGE.
Smith, J. A., Flowers, P., & Larkin, M. (2009). *Interpretative phenomenological analysis: theory, method and research*. London;Los Angeles, [Calif.];: SAGE.
Smith, J. A., Jarman, M., & Osborn, M. (1999). Doing interpretative phenomenological analysis. In *Qualitative health psychology: Theories & methods* (pp. 218-240).
Smith, J. A., & Osborn, M. (2008). Interpretative phenomenological analysis. In J. A. Smith (Ed.), *Qualitative psychology: a practical guide to research methods* (2nd ed., pp. 53 - 80). London;Los Angeles, Calif;: SAGE.
Spolsky, B. (2008). Introduction:Language Testing at 25: Maturity and responsibility? *, 25*(3), 297-305. doi:10.1177/0265532208090153

Spolsky, B. (2017). History of language testing. In E. Shohamy, L. G. Or, & S. May (Eds.), *Language testing and assessment* (3 ed., pp. 375-384): Springer.
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.
Suto, I. (2012). A Critical Review of Some Qualitative Research Methods Used to Explore Rater Cognition. *Educational Measurement: Issues and Practice, 31*(3), 21-30. doi:10.1111/j.1745-3992.2012.00240.x
Tavakoli, P. (2011). Pausing patterns: Differences between L2 learners and native speakers. *ELT Journal, 65*(1), 71-79.
Tavakoli, P., Nakatsuhara, F., & Hunter, A. M. (2020). Aspects of Fluency Across Assessed Levels of Speaking Proficiency. *The Modern Language Journal, 104*(1), 169-191. doi:10.1111/modl.12620
Thomson, P. (2017). Vignette variations. Retrieved from https://patthomson.net/2017/05/29/variations-to-the-journal-article-genre/
Thorndike, E. (1910). Handwriting. Part I. The measurement of the quality of handwriting: The derivation of the scales. In (Vol. 11, pp. 67-69). Teachers College Record.
Training, M. o. E. a. (2015). *Ban hành định dạng đề thi đánh giá năng lực sử dụng tiếng Anh từ bậc 3 đến bậc 5 theo khung năng lực ngoại ngữ 6 bậc dùng cho Việt Nam, Hà Nội, ngày 11 tháng 3 năm 2015*. Vietnam
Tran, T. H. P., Nguyen, H., Do, T. T. H., Dang, T., Huynh, A. T., & Davidson, F. (2015). *A validation study on the newly-developed Vietnamese Standardised Test of English Proficiency (VSTEP)*. Paper presented at the Language Testing Research Colloquium, Toronto, Canada.
Treffers-Daller, J., Parslow, P., & Williams, S. (2018). Back to Basics: How Measures of Lexical Diversity Can Help Discriminate between CEFR Levels. *Applied Linguistics, 39*(3), 302-327. doi:10.1093/applin/amw009
Uchihara, T., & Clenton, J. (2020). Investigating the role of vocabulary size in second language speaking ability. *Language teaching research : LTR, 24*(4), 540-556. doi:10.1177/1362168818799371
Ure, J. (1971). Lexical density and register differentiation. *Applications of linguistics, 443452*.
Vaughan, C. (1992). Holistic assessment: what goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111-126). Norwood: NJ: Ablex Publishing Corp.
Vietnam, G. (2008). *Government decision 1400: Teaching and learning foreign languages in the national educational system, period 2008-2020*. Hanoi: VN:Author
Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197-223.
Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287.
Weigle, S. C. (1999). Investigating Rater/Prompt Interactions in writing assessment: Quantitative and Qualitative approaches. *Assessing Writing, 6*(2), 145-178.
Weigle, S. C. (2002). *Assessing Writing*: Cambridge University Press.
Weir, C. J. (2019). Global, local or glocal: Alternative pathways in English language test provision. In L. I. W. Su, C. J. Weir, & J. R. W. Wu (Eds.), *English Language Proficiency Testing in Asia: A New Paradigm Bridging Global and Local Contexts*: Taylor & Francis.
Wenger, E., McDermott, R. A., & Snyder, W. (2002). *Cultivating communities of practice: a guide to managing knowledge*. Boston, Mass: Harvard Business School Press.
Wertz, F. J. (2005). Phenomenological Research Methods for Counseling Psychology. *Journal of Counseling Psychology, 52*(2), 167-177. doi:10.1037/0022-0167.52.2.167

Wicaksono, R. (2020). Native and Non-native Speakers of English in TESOL. In *Ontologies of English* (pp. 80-98).
Widdowson, H. G. (1983). *Learning purpose and language use*: Oxford University Press.
Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing, 14*(1), 85-106.
Willey, K., & Gardner, A. (2011). *Building a community of practice to improve inter marker standardisation and consistency*.
Willis, J. W., Jost, M., & Nilakanta, R. (2007). *Foundations of qualitative research: interpretive and critical approaches*. London: SAGE.
Winke, P. (2012). Rating oral language. *The encyclopedia of applied linguistics*, 1-8.
Winke, P., Gass, S., & Myford, C. (2011). The Relationship Between Raters' Prior Language Study and the Evaluation of Foreign Language Speech Samples. *ETS Research Report Series, 2011*(2), i-67. doi:10.1002/j.2333-8504.2011.tb02266.x
Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing, 30*(2), 231-252. doi:10.1177/0265532212456968
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. doi: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. doi: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. doi:10.1111/j.1745-3992.2012.00241.x
Wu, J. (2014). *Ensuring quality and fairness in the Asian EFL context: Challenges and opportunities*. Paper presented at the The 5th ALTE International Conference, Paris, France.
Wu, J. (2016). *A locally appropriate English language test - locality, globality & validity*. Paper presented at the The 4th British Council New Directions in English Language Assessment, Hanoi, Vietnam.
Xi, X., & Mollaun, P. (2009). *How do raters from India perform in scoring the TOEFL iBT speaking section and what kind of traning helps?* Retrieved from Princeton, New Jersey:
Xi, X., & Mollaun, P. (2011). Using Raters From India to Score a Large-Scale Speaking Test. *Language Learning, 61*(4), 1222-1255. doi:10.1111/j.1467-9922.2011.00667.x
Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *Language Testing, 31*(4), 501-527. doi:10.1177/0265532214536171
Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing, 28*(1), 31-50. doi:10.1177/0265532209360671

Appendix 1 – Sample of a VSTEP speaking test

The speaking test consists of three parts (see Appendix A for a sample of a test). In Part 1 (Social interaction), the test-takers are required to answer 3-6 questions on two different topics. This part lasts about 3-4 minutes. In Part 2 (Situation), the test-taker are given a situation with three options to select. The test-taker are required to select the best option and explain the reason(s) why that option is chosen and why the other two are not. The test-taker have 3-4 minutes for their explanation of their choice. The third part in the test, which lasts 3-4 minutes, provides the test-taker with a topic and a mind map of suggested ideas of how to develop the given topic. The test-taker can use the suggested ideas or his/her own ideas. If time allows, the test-taker discuss several follow-up questions upon the completion of Part 3 (Topic development).

PAPER 4. SPEAKING

Time allowance: 12 minutes
Number of questions: 3

Part 1: Social Interaction (3')

Let's talk about your free time activities.

- What do you often do in your free time?
- Do you watch TV? If no, why not? If yes, which TV channel do you like best? Why?
- Do you read books? If no, why not? If yes, what kinds of books do you like best? Why?

Let's talk about your neighborhood.

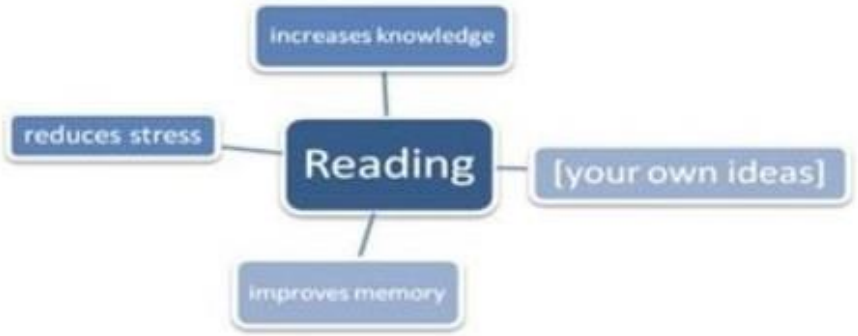
- Can you tell me something about your neighborhood?
- What do you like most about it?
- Do you plan to live there for a long time? Why/why not?

Part 2: Solution Discussion (4')

Situation: A group of people is planning a trip from Danang to Hanoi. Three means of transport are suggested: by train, by plane, and by coach. Which means of transport do you think is the best choice?

Part 3: Topic Development (5')

Topic: Reading habit should be encouraged among teenagers.



- *What is the difference between the kinds of books read by your parents' generation and those read by your generation?*
- *Do you think that governments should support free books for all people?*
- *In what way can parents help children develop their interest in reading?*

Appendix 2 – Sample of VSTEP speaking rating scale

	0	1	2	3
Grammar - Range - Accuracy	Test-taker is not present.	Performance does not satisfy Band 2 descriptors	Shows only limited control of a few simple grammatical structures and sentence patterns in a learnt repertoire	Uses some simple structures correctly, but still systematically makes basic mistakes; however, he/she can manage to make himself/herself understood
Vocabulary - Range - Control	Test-taker is not present.	Performance does not satisfy Band 2 descriptors	Only uses a basic vocabulary repertoire of isolated words and phrases related to particular concrete topics	- Uses appropriate vocabulary and can control a narrow repertoire dealing with familiar situations
Pronunciation - Individual sounds - Stress - Intonation	Test-taker is not present.	Performance does not satisfy Band 2 descriptors	- Is often unintelligible - Can articulate a very limited repertoire of learnt words and phrases with limited accuracy	- Is mostly intelligible - Can articulate simple words and phrases but conversational partners will need to ask for repetition from time to time

Fluency - Hesitation - Extended speech	Test-taker is not present.	Performance does not satisfy Band 2 descriptors	Can only manage very short, isolated words and phrases, mainly learnt utterances, with much pausing	Can construct short words and phrases with noticeable hesitation, frequent false starts, and repetition
Discourse Management - Thematic development - Coherence and cohesion	Test-taker is not present.	Performance does not satisfy Band 2 descriptors	- Hardly expresses or develop his/her ideas - Only links words or groups of words with very basic connectors like 'and' or 'then'	- Expresses his/her ideas with limited relevance to questions and cannot develop ideas without relying heavily on the repetition of the prompts - Links groups of words with simple connectors like 'and', 'but' and 'because'

Appendix 3 – Conference papers and awards

Awards & Grants

Santander Research Support Fund, for participating in LTRC2021 pre-workshop and presenting paper at the conference, awarded by the Graduate School, University of Huddersfield, June 2021

The Research Assessment Award – British Council, April 2019, for the current PhD research project at University of Huddersfield, the UK, awarded by British Council, 2019, valued at 2,050 GBP. This award requires the submission of two peer-reviewed reports. The information of the award recipients can be found at Assessment Research Awards | British Council

Best Student Paper, "Voices from teacher-raters in scoring speaking performances in a high-stakes localised test of English proficiency", presented at the 6th annual conference of Asian Association of Language Testing in Hanoi, Vietnam, October 2019. The content of the paper was

part of the PhD project and the travel expenses were funded by the School of Education and Professional Development, University of Huddersfield

Best Poster Prize, "Teacher-raters' beliefs and their practices in scoring a high-stakes test", presented at the annual conference - Language Testing Forum, UK Association of Language Testing and Assessment, in University of Bedfordshire, the UK, November 2018. The content of the paper was part of the PhD project and the travel expenses were funded by the School of Education and Professional Development, University of Huddersfield

Selected Conference Presentations

Factors influencing raters' scoring decisions: A study of a high-stakes test in Vietnam

Presented at the 42nd Language Testing Research Colloquium, organised by International Language Testing Association, Virtual on 14-17 June 2021

Voices from teacher-raters in scoring speaking performances in a high-stakes localised test of English proficiency

Presented at the 6th annual conference of Asian Association of Language Testing in Hanoi, Vietnam, October 2019

Teacher-raters' beliefs and their practices in scoring a high-stakes test

Presented at the annual conference - Language Testing Forum, UK Association of Language Testing and Assessment, in University of Bedfordshire, UK, November 2018

Factors influencing raters' scoring decisions and their expertise development: A study of a high-stakes test in Vietnam (A research proposal)

Presented at the annual conference of the School of Education and Professional Development, University of Huddersfield, April 2018

Appendix 4 – Participation information sheet & Consent form

University of HUDDERSFIELD
Inspiring tomorrow's professionals

2013 THE AWARDS
AWARD WINNER
UNIVERSITY OF THE YEAR

PARTICIPATION INFORMATION SHEET AND INFORMED CONSENT FORM

1. **Invitation to participate**:
   You are being invited to take part in my PhD research. Before you decide it is important for you to understand why this research is being done and what it will involve. Please take time to read the following information and discuss it with others if you wish. Ask if there is anything that is not clear or if you would like more information.
2. **British Educational Research Association (BERA) ethical guidelines**
   This research will be carried out in line with BERA's 2018 guidelines for educational research. I am happy to provide you with a copy of these guidelines if you wish to read them before agreeing to participate in the research
3. **The research project and its title**
   The aim of the research is to explore factors that may influence raters' scoring decision and their expertise development. My working title for the research is 'Factors influencing raters' scoring decision and their expertise development: A case study of a high-stakes proficiency test in Vietnam'.
4. **What is the purpose of the project?**
   I am studying for PhD degree and the research is being completed as part of my study requirement.
5. **Why have I been chosen?**
   I have approached you as you have experience of rating VSTEP speaking tests, so I think you will be best placed to inform this research.
6. **Do I have to participate?**
   No. Your participation is entirely voluntary and you may withdraw from the research at any time, and for any reason. If you feel unable to be involved for any reason, I fully understand.
7. **What do I have to do?**
   You will complete the questionnaire I have distributed to you. I plan to invite you to participate in one rating moderation before you mark the responses. I will then invite you to do the verbal protocol while rating speaking performances. I plan to tape your speaking out of your ratings. I plan to follow this up with a semi-structured interview to further discuss your rating experience. I plan to tape your interviews.
8. **Are there any disadvantages to taking part?**
   I foresee no disadvantages to participating in this study. However, you may feel mentally and/or physically fatigued from the length of time that is required for you perform the ratings. To minimise the fatigue, you are welcome to stretch, get a snack, or use the restroom at your discretion during your rating sessions.
   You also may feel uncomfortable with someone analysing your ratings. The purpose of the study is not to draw conclusions about your effectiveness as a rater, but rather to gather information about the scoring experience.

9. **Will all my details be kept confidential?**
   In line with the Data Protection Act, the consent form, questionnaire data, taped moderation, taped rating process and taped interviews will be securely stored by me during the research. You may access the material I collect from you at any time during the research. To ensure your anonymity, I will ask you to choose a pseudonym so that when I make any reference to you in the research your identity will be protected.
10. **What will happen to the results of the research project?**
    I will write up the research and it will be presented to meet the assessment requirement of my PhD study. I will securely dispose of the recordings, interview tapes and my research notes after the conclusion of the research.

Consent:

I have read, understood, and received a copy of the above consent and desire of my own free will to participate in this study.

I agree that my contribution including verbatim quotations may be used, as long as it protects anonymity/confidentiality

Name: Signature:

Date:

Name of researcher: Signature:

Date:

Contact address:

Thuy Thai
School of Education and Professional Development,
University of Huddersfield,
Queensgate, Huddersfield.
HD1 3DH
Email: { HYPERLINK "mailto:thuy.thai@hud.ac.uk" }

Appendix 5 – Rater's briefing sheet

Raters' briefing sheet for a spoken protocol of the marking process

The main aim of this exercise is to obtain and record your authentic response to VSTEP speaking performances. I would like to find out the marking strategies used by raters. The only way to be sure of eliciting this information is to tape you thinking aloud as you mark. What I mean by 'thinking aloud' is that I want you to tell me EVERYTHING you are thinking from the time you first listen to each performance until you finish your rating procedure.

I would be grateful if you can follow the instructions below:

Before you start

Approach the recorded performances in the same way you often do with any of the other performances you are marking.

Start the tape recorder running just before you start listening to the script

Please don't switch it off before you have finished marking the script, even if you have long periods where you cannot think of anything to say, or where you are pausing over a section. It is important for me to know how long each script, and each section of a script, demands your attention.

During your marking

Start by reading out the test-taker's number before you begin the marking. This will enable me to match your verbal protocol with the script.

As you mark, try literally to speak aloud your thoughts on the script, as much as possible without censoring or editing them. This will enable me to know about the strategies you often use while marking.

If you would normally pause and take notes of a script, please do so but I would appreciate if you could

Read aloud the comment as you write it

Refer to other material (rating scales, phrases, words) if you do so

Refer to the rough time (indicated in the cassette)

Repeat or describe the section that attracts your attention

You can listen to part of the performance or the whole performance again, and please let me know which part and why since it is really helpful to my study.

After you mark

Please try to summarise all the factors that have helped you reach that mark, whatever they are.

It may seem from these instructions that it will take a very long time, but I would estimate that it shouldn't take more than double the time of your typical marking.

Appendix 6 – Interview questions

Warm-up

What do you think about VSTEP test?

How did you become a VSTEP rater?/ Why did you decide to become a rater?

How long have you worked as a VSTEP rater?

Do you work as a rater for other tests?

How often do you participate in VSTEP rating process?

What advantages does being a rater bring to you? (personal, teaching, learning)

Can you share with me what challenge you as a rater?

Rating process / strategies and expertise development

1. Can you talk me through a typical speaking test?

When (roughly) was the first time you marked a VSTEP speaking test?

Can you describe the very first rating process you did when you first became a rater?

Did you encounter any difficulties in rating when you first became a VSTEP rater? What were they?

What did you do/ have you done to minimise the difficulties?

When was your last time rating a VSTEP speaking performance?

Can you talk me through your last rating process? (Prompt when scores are awarded)

What do you consider important in rating a VSTEP speaking performance? Please explain your reasons.

2. Can you describe how you use the rating scales?

What criteria do you start your ratings with?

Can you put in order all the criteria from easiest to most difficult to mark? Please explain your reasons.

Fluency Grammar Vocabular Pronunciation Discourse management

3. When you need to make decision for borderline performances, what do you consider important? Please explain why.

4. Can you describe situations/performances in which you feel it is hard to make scoring decisions?

What do you do in those situations?

5. Have you noticed any differences in your ratings over time?

Can you tell me what they are?

What might contribute to the differences?

Have you received information/feedback about your ratings? (quick, consistent, etc.)

What do you think about the information/feedback?

Score-influencing factors

6. What do you think about when you are scoring a speaking performance?

What do you consider important in scoring a speaking performance? Can you explain why?

7. Can you tell me about your English learning experience?

How did you learn English, particularly speaking skills?

Which English standard did you learn?

Which English standard do you follow?

Do you have any preference toward a particular standard of English?

Do you know how people (colleagues, friends, students, teachers, etc.) think about your English speaking?

8. Can you tell me about your teaching experience?

How long have you been teaching English?

Can you describe your students' characteristics? (age group, learning needs, learning targets,

How important is speaking skills to your students?

What do you consider important in teaching speaking skills to your students? Can you list them according to importance levels? Please explain why.

How do you consider errors/mistakes in learning speaking?

How do you address them in your teaching?

Which English standard do you teach your students?

What is the role of teaching experience in your ratings?

- Do you often refer back to those experiences when scoring the performance?

9. How many trainings have you attended in terms of language testing and assessment?

What did you do in the training?

What is the role of such training in your experience of ratings?

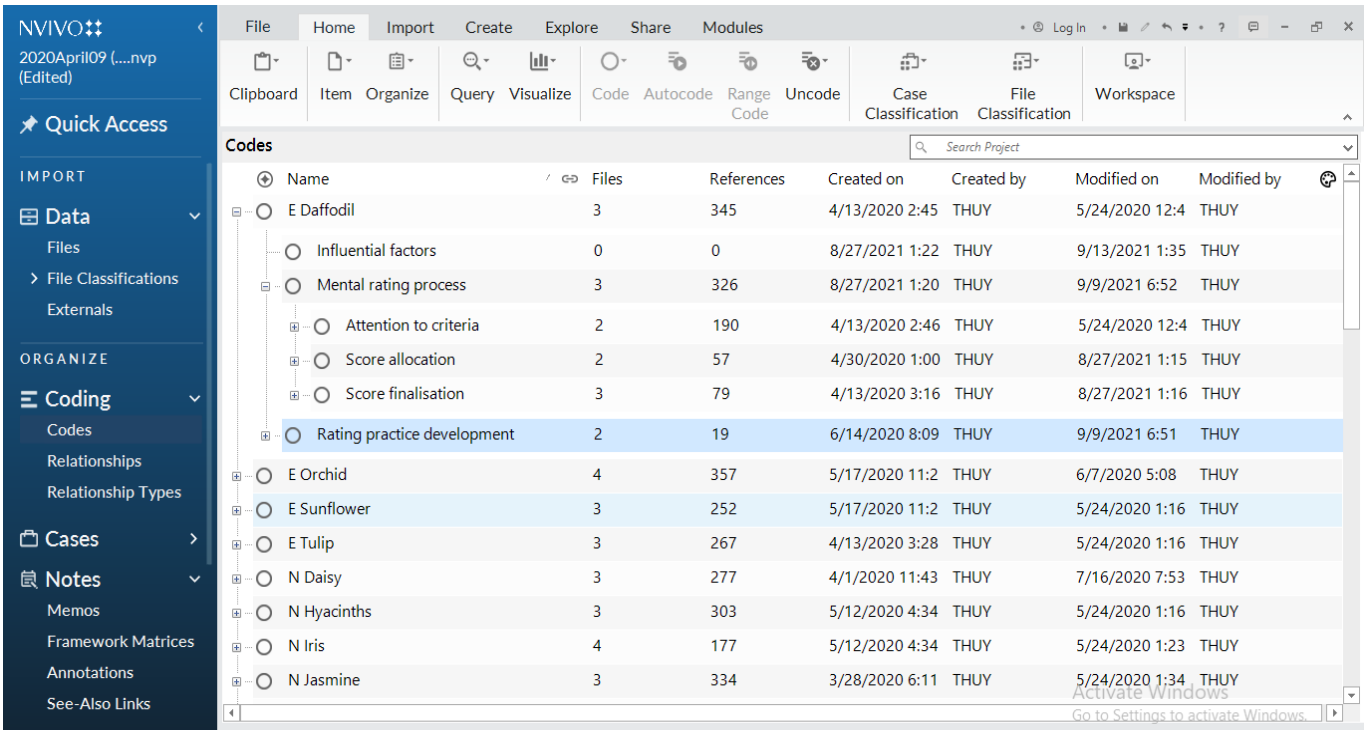
10. Can you share with me in your opinion what make a good rater?

Can you suggest what can be done to become a good rater? (individual level, organisation level, etc.)

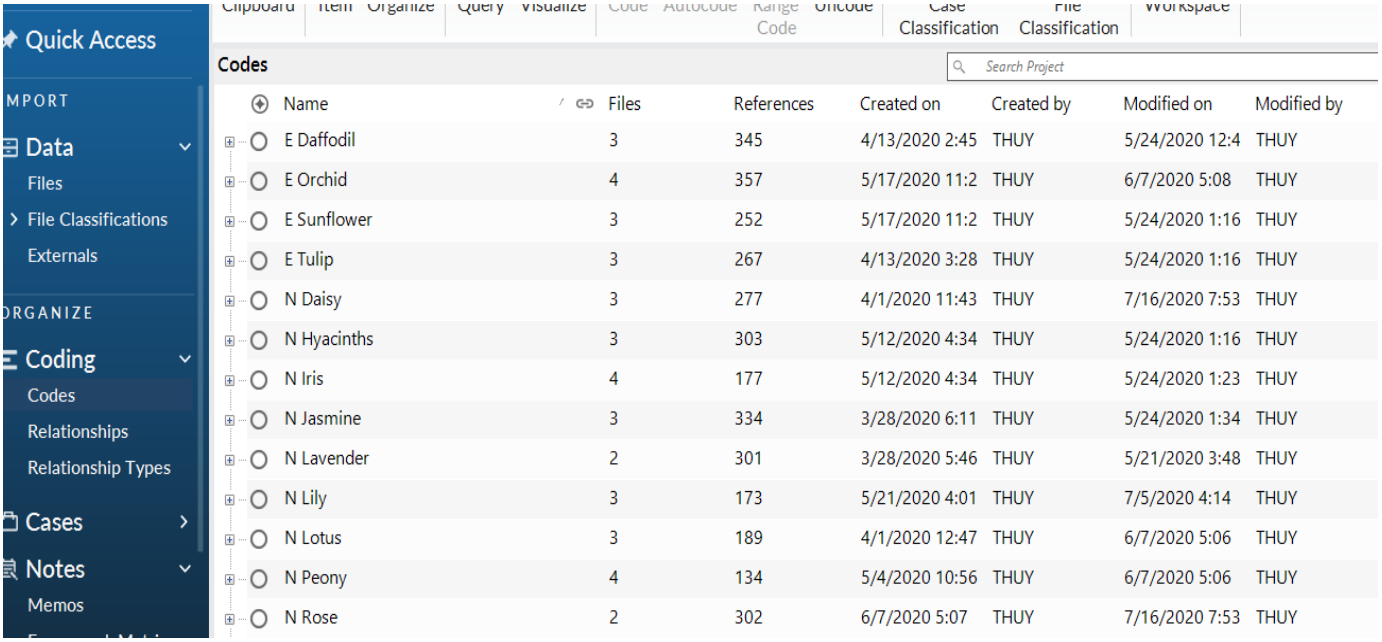
Knowledge

Skills

Appendix 7 – A template of coding

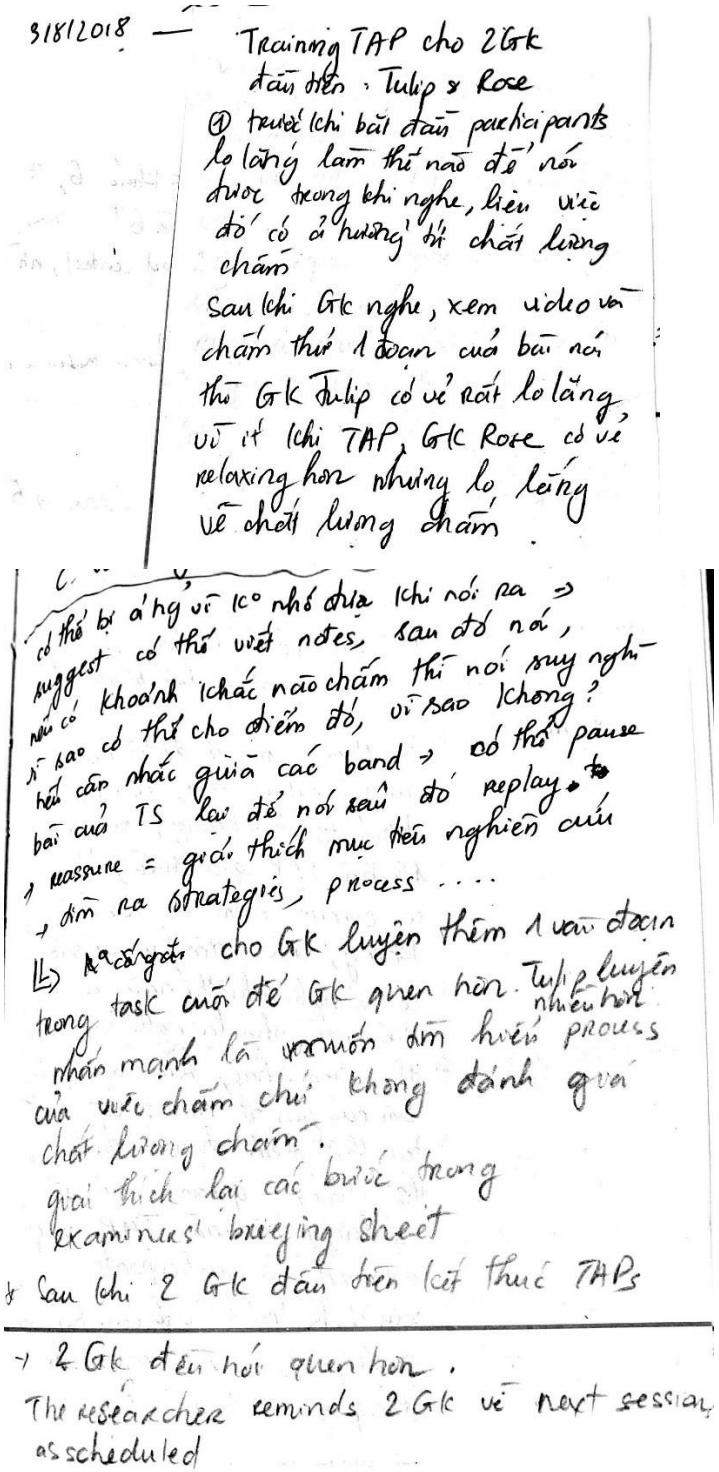


Name	Files	References	Created on	Created by	Modified on	Modified by
E Daffodil	3	345	4/13/2020 2:45	THUY	5/24/2020 12:4	THUY
Influential factors	0	0	8/27/2021 1:22	THUY	9/13/2021 1:35	THUY
Mental rating process	3	326	8/27/2021 1:20	THUY	9/9/2021 6:52	THUY
Attention to criteria	2	190	4/13/2020 2:46	THUY	5/24/2020 12:4	THUY
Score allocation	2	57	4/30/2020 1:00	THUY	8/27/2021 1:15	THUY
Score finalisation	3	79	4/13/2020 3:16	THUY	8/27/2021 1:16	THUY
Rating practice development	2	19	6/14/2020 8:09	THUY	9/9/2021 6:51	THUY
E Orchid	4	357	5/17/2020 11:2	THUY	6/7/2020 5:08	THUY
E Sunflower	3	252	5/17/2020 11:2	THUY	5/24/2020 1:16	THUY
E Tulip	3	267	4/13/2020 3:28	THUY	5/24/2020 1:16	THUY
N Daisy	3	277	4/1/2020 11:43	THUY	7/16/2020 7:53	THUY
N Hyacinths	3	303	5/12/2020 4:34	THUY	5/24/2020 1:16	THUY
N Iris	4	177	5/12/2020 4:34	THUY	5/24/2020 1:23	THUY
N Jasmine	3	334	3/28/2020 6:11	THUY	5/24/2020 1:34	THUY



Name	Files	References	Created on	Created by	Modified on	Modified by
E Daffodil	3	345	4/13/2020 2:45	THUY	5/24/2020 12:4	THUY
E Orchid	4	357	5/17/2020 11:2	THUY	6/7/2020 5:08	THUY
E Sunflower	3	252	5/17/2020 11:2	THUY	5/24/2020 1:16	THUY
E Tulip	3	267	4/13/2020 3:28	THUY	5/24/2020 1:16	THUY
N Daisy	3	277	4/1/2020 11:43	THUY	7/16/2020 7:53	THUY
N Hyacinths	3	303	5/12/2020 4:34	THUY	5/24/2020 1:16	THUY
N Iris	4	177	5/12/2020 4:34	THUY	5/24/2020 1:23	THUY
N Jasmine	3	334	3/28/2020 6:11	THUY	5/24/2020 1:34	THUY
N Lavender	2	301	3/28/2020 5:46	THUY	5/21/2020 3:48	THUY
N Lily	3	173	5/21/2020 4:01	THUY	7/5/2020 4:14	THUY
N Lotus	3	189	4/1/2020 12:47	THUY	6/7/2020 5:06	THUY
N Peony	4	134	5/4/2020 10:56	THUY	6/7/2020 5:06	THUY
N Rose	2	302	6/7/2020 5:07	THUY	7/16/2020 7:53	THUY

Appendix 8 – Sample of research journal



3/8/2018 — Training TAP cho 2 Gk đầu tiên: Tulip & Rose

① trước khi bắt đầu participants lo lắng làm thế nào để nói được trong khi nghe, liệu việc đó có ảnh hưởng tới chất lượng chấm

Sau khi Gk nghe, xem video và chấm thử 1 đoạn của bài nói thì Gk Tulip có vẻ rất lo lắng, vì ít khi TAP, Gk Rose có vẻ relaxing hơn nhưng lo lắng về chất lượng chấm.

có thể bị ảnh với Ic° nhớ chưa khi nói ra ⇒ suggest có thể viết notes, sau đó nói, nếu có khoảnh khắc nào chấm thì nói suy nghĩ xem sao có thể cho điểm đó, vì sao không? nếu cân nhắc giữa các band ⇒ có thể pause bài của TS lại để nói sau đó replay lại

↳ reassure = giải thích mục tiêu nghiên cứu

↳ đưa ra strategies, process ....

↳ đề xuất cho Gk luyện thêm 1 vài đoạn trong task cuối để Gk quen hơn. Tulip luyện nhiều hơn

nhấn mạnh là muốn đưa thêm process của việc chấm chứ không đánh giá chất lượng chấm.

giải thích lại các bước trong examiners' briefing sheet

* Sau khi 2 Gk đầu tiên kết thúc TAPs

→ 2 Gk đến nói quen hơn.

The researcher reminds 2 Gk về next session as scheduled

Translation:

3/8/2018 – Training TAP for the first two raters: Tulip and Rose

1. Before the training, the participants seem to be worried of how they can speak while listening to the recorded speaking performances and if that will affect their rating quality

After the raters listened, watched videos and rated one part of the recorded performance, Tulip seemed to be very nervous because she rarely does the TAPs while Rose seemed to be more relaxing but worried that the rating quality might be affected because she thought she could not retain the information while talking out loud. I suggested that she could write notes down, then talk outloud the notes. Then when she awarded the score, talked out loud her thoughts why she decided that score, why not different scores. If she considered different band scores, she could pause the recording to talk out loud why and then replay the recording. I reassured the participants by explaining to them again the aims of the study, that is, to understand the

818

Appointment with:

Dayodil, Lily, Jasmine } team TAP + TAP$_1$ (6)

Tulip, Rose } TAP 2

Jasmine reported a technical problem with the CD as she could not hear clearly the last part of performance 4. Luckily I had prepared more than CDs needed. I gave her another one & it worked well.

After TAP$_1$, Dayodil told me that she'd love to have some feedback on her rating pattern or if she could change anything related to her ratings. I thought she might have thought I came from the centre & did this to judge her rating. I gently told her my study's focus is on the rating process & features & that if she need to change anything, the team leader would have told her. So "please don't think that I will judge your scoring quality. But I will be happy to share with you my findings when it's available." She told me she was not afraid of being judged, just curious of her rating patterns & she's interested in the findings of my study.