THE REPUBLIC OF TURKEY
ÇANAKKALE ONSEKİZ MART UNIVERSITY
SCHOOL OF GRADUATE STUDIES

DEPARTMENT OF FOREIGN LANGUAGES EDUCATION
ENGLISH LANGUAGE TEACHING PROGRAM

THE EFFECT OF RATER EXPERIENCE AND L2 SPEAKING PERFORMANCE QUALITY ON SCORE VARIATION AND RATER BEHAVIOR

DOCTORAL DISSERTATION

MUSTAFA ÇOBAN

SUPERVISOR
ASSOC. PROF. DR. SALİM RAZI

ÇANAKKALE – 2022

THE REPUBLIC OF TURKEY
ÇANAKKALE ONSEKİZ MART UNIVERSITY
SCHOOL OF GRADUATE STUDIES

DEPARTMENT OF FOREIGN LANGUAGES EDUCATION
ENGLISH LANGUAGE TEACHING PROGRAM

THE EFFECT OF RATER EXPERINCE AND L2 SPEAKING PERFORMANCE QUALITY ON SCORE VARIATION AND RATER BEHAVIOR

DOCTORAL DISSERTATION

MUSTAFA ÇOBAN

Supervisor
ASSOC. PROF. DR. SALİM RAZI

This dissertation was supported by TOEFL ETS Small Grants for Doctoral Research in Second or Foreign Language Assessment

ÇANAKKALE – 2022

T.C.
ÇANAKKALE ONSEKİZ MART ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

Mustafa ÇOBAN tarafından Doç. Dr. Salim RAZI yönetiminde hazırlanan ve **18/07/2022** tarihinde aşağıdaki jüri karşısında sunulan **"Puanlayıcı tecrübesi ve ikinci dil konuşma performans kalitesinin puan değişkenliği ve puanlayıcı davranışı üzerindeki etkisi"** başlıklı çalışma, Çanakkale Onsekiz Mart Üniversitesi Lisansüstü Eğitim Enstitüsü **Yabancı Diller Eğitimi Anabilim Dalı**'nda **DOKTORA YETERLİK TEZİ** olarak oy birliği ile kabul edilmiştir.

Jüri Üyeleri	**İmza**
Doç. Dr. Salim RAZI (Danışman)	……
Prof. Dr. Aysun YAVUZ	……
Prof. Dr. A. Amanda J. A. YEŞİLBURSA	……
Doç. Dr. Serdar ARCAGÖK	……
Doç. Dr. Turgay HAN	……

Tez No : 10486644
Tez Savunma Tarihi : 18/07/2022

……
Doç. Dr. Yener PAZARCIK
Enstitü Müdürü
../../20

DECLARATION OF ETHICS

I hereby undertake and declare that in this doctoral dissertation, I prepared in accordance with the rules of dissertation writing procedures of School of Graduate Studies at Çanakkale Onsekiz Mart University, I have obtained the data, information and documents, presented in the dissertation within the framework of academic and ethical rules, I have presented all the information, documents, evaluations and results in accordance with the code of scientific ethics and ethics, all sources which I have utilized during the dissertation have been fully cited in the references, I have not made any changes to the data used that the study I have presented in this dissertation is original, which I have accepted all losses of rights that may arise against me otherwise.

ETİK BEYAN

Çanakkale Onsekiz Mart Üniversitesi Lisansüstü Eğitim Enstitüsü Tez Yazım Kuralları'na uygun olarak hazırladığım bu tez çalışmasında; tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi, tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu, tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi, kullanılan verilerde herhangi bir değişiklik yapmadığımı, bu tezde sunduğum çalışmanın özgün olduğunu, bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi taahhüt ve beyan ederim.

Mustafa ÇOBAN
18/07/2022

ACKNOWLEDGEMENT

This dissertation could not have been finalized without the guidance and support of many people. First and foremost, I would like to express my deepest gratitude to my supervisor Assoc. Prof. Dr. Salim Razı, who has always been generous, friendly, and patient while providing feedback. He has always inspired and encouraged me to be a true scholar throughout this long PhD journey. Secondly, I deeply appreciate Dr. Özgür Şahan for his guidance and advice at every process of writing this dissertation. Without his insightful feedback, I would not have made a steady progress in analyses. In addition, I would like to thank my dissertation committee members Prof. Dr. Aysun Yavuz, Prof. Dr. Ayşegül Amanda Yeşilbursa, Assoc. Prof. Dr. Serdar Arcagök, and Assoc. Prof. Dr. Turgay Han for their guidance and feedback.

My sincere thanks goes to all the students and raters who contributed to my dissertation as participants. I would like to thank my friend Dr. Kari Coffman Şahan. Her great support and guidance helped me progress in my qualitative data analysis. My sincere and deep gratitude goes to my beloved wife Betül. Without her support, understanding, and love, the completion of this dissertation would have been impossible. I would like to attribute this dissertation to many unplayed games and untold stories because my lovely sons Yekta, Olcay, and Görkem had mostly been looking forward to spending time with me. Without their love, everything would have been meaningless to me.

Lastly, this dissertation was funded by TOEFL ETS Small Grants for Doctoral Research in Second or Foreign Language Assessment in 2018. I gratefully acknowledge the financial support from this grant for enhancing the completion of my dissertation.

Çanakkale, 2022

Mustafa ÇOBAN

ABSTRACT

THE EFFECT OF RATER EXPERIENCE AND L2 SPEAKING PERFORMANCE QUALITY ON SCORE VARIATION AND RATER BEHAVIOR

Mustafa ÇOBAN

Çanakkale Onsekiz Mart University

School of Graduate Studies

Department of Foreign Languages Education

Doctoral Dissertation in English Language Teaching Program

Supervisor: Assoc. Prof. Dr. Salim RAZI

18/07/2022, 215

The purpose of this dissertation was to examine the effect of rater experience and L2 speaking performance quality on score variation and rater behavior. Utilizing convergent parallel case study mixed-method design, both quantitative and qualitative methodologies were combined to address the issues of rater experience and L2 speaking performance quality. Twenty-five EFL instructors participated in this study. All the participants were from the same context, a university in western Türkiye. Using a rater experience scale form, three rater experience groups were formed: low-experienced (n = 10), medium-experienced (n = 7), and high-experienced raters (n = 8). Using an analytic rubric, the participant raters (n = 25) evaluated a number of 60 three quality L2 speaking performances. They also completed verbal protocols as well as written score explanations, corroborating the results retrieved from the quantitative data. The participants provided 7,500 scores (1,500 total scores and 6,000 sub-scores), 375 verbal protocols and 4,500 written score explanations.

The results showed that the analytic scores assigned to low-quality, medium-quality, and high-quality L2 speaking performances were statistically significant different from each other. However, rater experience groups did not differ significantly in their both total and

component scores assigned to three speaking performance qualities. Furthermore, the results of Generalizability study (G-study) suggested that there was a limited rater impact on the variation when all L2 speaking performances were examined in total, yet more rater effect was observed when speaking performance qualities were analyzed individually. The qualitative findings revealed that raters showed certain decision-making behaviors across three speaking performance qualities.

Key Words: L2 Speaking Assessment, Generalizability Theory, Rater Behavior, Rater Experience, Score Variation, Verbal Protocols

ÖZET

PUANLAYICI TECRÜBESİ VE İKİNCİ DİL KONUŞMA PERFORMANS KALİTESİNİN PUAN DEĞİŞKENLİĞİ VE PUANLAYICI DAVRANIŞI ÜZERİNDEKİ ETKİSİ

Mustafa ÇOBAN

Çanakkale Onsekiz Mart Üniversitesi

Lisansüstü Eğitim Enstitüsü

Yabancı Diller Eğitimi Anabilim Dalı

(İngiliz Dili Eğitimi Programı Doktora Yeterlik Tezi)

Danışman: Doç. Dr. Salim RAZI

18/07/2022, 215

Bu tezin amacı, puanlayıcıların puanlama deneyimlerinin ve değerlendirilen konuşma sınavlarının kalitesinin puan değişkenliği ve puanlayıcı davranışı üzerindeki etkilerini incelemektir. Yakınsayan paralel tasarım, durum çalışması karma yöntem araştırma yaklaşımı kullanılarak, puanlayıcıların puanlama deneyiminden ve konuşma sınavlarındaki yanıtların kalitesinden kaynaklanan sorunları ele almak için hem nicel hem de nitel araştırma yöntemleri kullanıldı. Bu çalışmaya 25 İngilizce öğretim görevlisi konuşma sınavı puanlayıcısı olarak katılmıştır. Tüm katılımcılar aynı araştırma bağlamından olup, Türkiye'nin batısındaki bir üniversitede görev yapmıştır. Puanlayıcı deneyim ölçeği kullanılarak, düşük deneyimli grup 10, orta deneyimli grup 7 ve yüksek deneyimli puanlayıcı grup 8 katılımcıdan olmak üzere üç adet deneyim grubu oluşturulmuştur.

Bütünsel puanlama ölçeği kullanarak, 25 katılımcının hepsi farklı kaliteden oluşan 60 adet konuşma sınavı yanıtını değerlendirdi. Ayrıca, katılımcılar verdikleri puanlara gerekçe oluşturdukları yazılı puan açıklamalarının yanı sıra sesli düşünme protokollerini tamamladılar. Her iki veri toplama yöntemi de nicel verilerden elde edilen bulguları

doğrulamak için kullanılmıştır. Katılımcılar 7,500 adet konuşma sınavı puanı (1,500 toplam puan ve 6,000 alt puan), 375 adet sesli düşünme protokolü ve 4,500 adet yazılı puan açıklaması oluşturdu.

Çalışmanın sonuçları, düşük kaliteli, orta kaliteli ve yüksek kaliteli konuşma sınavı yanıtlarına verilen puanların birbirinden istatistiksel olarak anlamlı farklılıklar gösterdiğini göstermiştir. Ancak, puanlayıcı deneyim grupları, üç farklı kalitedeki yanıtlara verilen hem toplam hem de bileşen puanlarında anlamlı farklılıklar göstermedi. Genellenebilirlik kuramı bulguları, tüm yanıtlar toplamda incelendiğinde varyasyon üzerinde sınırlı puanlayıcı etkisinin olduğunu, ancak farklı kalitedeki yanıtlar ayrı ayrı analiz edildiğinde daha fazla puanlayıcı etkisinin gözlemlendiğini ortaya koymuştur. Nitel bulgular, puanlayıcıların üç farklı kalitedeki yanıtları değerlendirirken belirli karar verme davranışları sergilediğini ortaya koydu.

Anahtar Kelimeler: Genellenebilirlik Kuramı, İngilizce Konuşma Becerisi Değerlendirme, Puanlayıcı Davranışı, Puanlayıcı Deneyimi, Puan Değişkenliği, Sesli Düşünme Protokolü

TABLE OF CONTENTS

Page Number

CHAPTER II

LITERATURE REVIEW

CHAPTER III

METHODOLOGY

CHAPTER IV

RESULTS

CHAPTER V

DISCUSSION AND CONCLUSIONS

ABBREVIATIONS

CEFR	Common European Framework of Reference
CTT	Classical Test Theory
D-Study	Decision Study
EBB	Empirically derived, Binary-choice, Boundary-definition
EFL	English as a Foreign Language
ELT	English Language Teaching
EMI	English Medium Instruction
EPP	English Preparatory Program
ESL	English as a Second Language
ETS	Educational Testing Service
G-Study	Generalizability Study
G-Theory	Generalizability Theory
IRT	Item Response Theory
L1	First Language
L2	Second Language
NES	Native English Speakers
NNES	Non-native English Speakers
VPA	Verbal Protocol Analysis
YÖK	Turkish Council of Higher Education

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1. Introduction

Language tests are nearly in all aspects of people's lives from an ordinary classroom to a complex recruitment process. Therefore, testing learners' language skills and abilities is one of the mainstays of language teaching and learning. Appreciating the magnitude of language testing, it is crucial to minimize errors and discrepancies stemming from measurement tools (Douglas, 2010). In fact, the teaching and learning process without proper assessment principles would misguide all stakeholders. Given the fundamental place of tests in this journey, a reliable and valid assessment system can help learners to improve their motivation, make some necessary changes and most importantly prod them to take responsibility for their own learning, all of which refer to an independent learner profile (Brown, 2004). Accordingly, it is of central importance to decide the goal of tests and consider their effects on learners and institutions, which is closely related to the development of a reliable, valid and fair assessment system (Bachman, 1990).

The inextricable connection between reliability and validity provides a link between language assessment and stakeholders in the cycle of testing. This connection informs us about the significance of building both a reliable and valid assessment system as language tests can literally either close or open a door for a candidate taking a high stake test (Fulcher, 2010). To prepare reliable and valid test items, task and context are two essential elements. Stressing the importance of determining the goal and need of tasks as well as the boundaries of contexts in speaking assessment, Luoma (2004) defines speaking tasks as "activities that involve speakers in using language for the purpose of achieving a particular goal or objective in a particular speaking situation" (p. 31). Indeed, tasks are not independent from objectives and contexts in speaking assessment. Therefore, designing speaking tasks will be more challenging than writing or any other test tasks since various factors are naturally embedded in all phases of the assessment (Hughes, 2010).

The assumption that the subjectivity of raters can be minimized using a set of standards or rubric criteria might not guarantee the reliability of assigned scores. Even if two raters with the same level of experience award similar scores, they may interpret them differently (Douglas, 1997). It would seem that prioritizing the overall performance of test takers instead of interpreting only the result of their performance is of primary importance because the outcomes of speaking tasks might be uncertain and difficult to interpret (Douglas, 2010). Given the complex nature of L2 speaking assessment, it can be claimed that numerous factors such as raters' characteristics, rating scales, speaking tasks, and test takers may influence the assigned scores, namely, the reliability of an assigned score depends on these interacting factors (Fulcher, 2014; Luoma, 2004; McNamara, 1996; Nunan, 1989). Considering humans' fundamental nature, raters' subjectivity could be listed as one of the major sources of score variations in performance assessment (Bachman, 2004; Brown, 2004; McNamara, 2000). This situation does not underestimate the value of human raters. On the contrary, it places them in the center of the rating process (Green, 2014). In fact, human raters have still been salient in terms of benchmarking the effectiveness of various automated rating systems (Isaacs, 2016).

Speaking assessment is a multicomponent field in which there are various factors affecting one another, and indeed, the ways how and reasons why speaking raters give their rating decisions seem to be a useful contribution to understanding reliability and fairness issues of assigned scores (Fulcher, 2003; Galaczi, & Taylor, 2018; Luoma, 2004; O'Sullivan, 2013). Resolving these issues is not an easy task for stakeholders in speaking assessment, specifically the scores assigned by different groups of raters and varying level of L2 speaking performances (Bonk & Ockey, 2003; Davis, 2016; Kim, 2015). Although it is challenging to understand the issues of reliability and fairness of given scores, it is crucial that test designers always follow the correct procedure since the impact of these exams on test takers' life might be critical (Bachman, 1990; Fulcher & Davidson, 2007; Henning, 1987; Hughes, 2003; Luoma, 2004).

Determining how and why speaking raters award scores is of critical importance to ensuring the reliability and consistency of assigned scores. While doing so, the areas that contribute to score variation such as rater characteristics, rater training, and rater experience

should be investigated in detail (Brown, 1995; Davis, 2016; Kim, 2015; Xi & Mollaun, 2011; Yan, 2014; Zhang & Elder, 2011). In addition to the issue of reliability and consistency of scores, it is vital to reveal speaking raters' decision making patterns and rating approaches in specific contexts since they could give new insights into building an institutional speaking assessment model (Ang-Aw & Goh, 2011; Pollitt & Murray, 1996). Therefore, a better understanding of reliability of assigned scores and raters' decision making strategies can lead to developing best practice models in speaking assessment.

1.2. Problem Statement

The increasing impact of internationalization has resulted in a change in educational assessment trends in higher education institutes with specific emphasis on second language (L2) performance assessment. Given the continuous efforts of the Turkish Council of Higher Education (YÖK) to follow new approaches in language education and become a competitive country on a global scale, English Preparatory Programs (EPP) have been offering intensive English instruction to the first year university students prior to their departmental studies. In addition to EPP's instruction role, they serve the purpose of assessing students' level of English, and to that end conduct not only proficiency tests, progress tests, achievement tests, diagnostic tests, and placement tests but also exchange program exams. However, due to the lack of standardization for assessing students' language skills, providing students with fair judgements may remain a problematic issue in practice. Therefore, assessing English as a Foreign Language (EFL) speaking performance is essential for both high-stakes and low-stakes tests conducted at Turkish universities.

Not having any standardized test applications, most of the EPPs might be faced with widespread criticisms of the reliability and validity of performance based tests that they have conducted. These issues are mostly attributed to the testing units, the main role of which is to prepare high- and low-stakes exams of preparatory programs. However, considering the workload, lack of enough staff and the number of students, most testing units do not have adequate time and capacity to show a considerable improvement in areas such as norming sessions, rater training, rubric development and most importantly creating test specifications. Considering the standardization of speaking performance assessment in the EFL context, it

would not be wrong to claim that there are varied speaking assessment protocols across institutions. In fact, while some of them implement a single rater system without a proper rubric or scale, others follow a certain protocol including inter-rater reliability and a valid rubric. What is more, testing conditions are the main motive for unreliable judgment in that some institutions may not have any speaking assessment policy. In such cases, the assigned scores at such institutions reflect personal impressions or inner thoughts. It seems that these ineffective speaking assessment practices are widespread across universities in Türkiye. Consequently, the availability of a speaking assessment policy that touches upon score variation is crucial for the purposes of establishing a fair testing system.

As for the nature of performance assessment, L2 speaking assessment includes numerous components such as raters, scales, performance, test takers, individual characteristics, training quality, and assessment philosophy, each of which is closely interrelated and is to be evaluated thoroughly. These components are most likely to have an effect on the assessment quality. However, raters in particular play a central role in rating processes, training sessions, and tests constructs. Therefore, neglecting the phenomenon of rater effect is likely to make assigned scores be far from test realities (Fulcher, 2015). In fact, rater variation is one of the most intriguing aspects of L2 speaking assessment (Bejar, 1985; Douglas, 1997; Fulcher, 2014; Ginther, 2013; Goh & Burns, 2012; Knight, 1992; Lumley & McNamara, 1995; Luoma, 2004; McNamara, 1996). More specifically, different patterns that raters follow while rating L2 speaking performances can be related to professional experience, educational background, linguistic background and the extent of leniency and severity of scoring (Davis, 2016; Hubbard et al., 2006; Isaacs & Thomson 2013; Winke et al., 2012). In addition to the aforementioned factors, raters' fatigue, rater training and task types can affect the score variation (Brown, 2003; Knight, 1992; Ling, Mollaun & Xi, 2014).

One of the most serious threats to building a reliable and valid speaking test is mostly posed by rater types who consciously or unconsciously grade test takers either too strictly or generously (Bejar, 2012). In line with this problem, raters' professional experience including the period of time spent on teaching and assessing L2 speaking might have an effect on the score variation (Bejar, 1985; Bonk & Ockey, 2003; Brown et al., 2005; Cai, 2015; Davis, 2016; Huang, 2013; Isaacs, & Thomson, 2013; Kuiken & Vedder, 2014). Thus, it is essential

to investigate the effect of rater experience on score variation and rater characteristics to address the aforementioned issues related to speaking performance assessment at higher education institutes in various contexts. In addition to rater experience, two points should be examined to get a more complete picture: a) how raters assess L2 speaking performances in different qualities (low, medium, and high) b) how varying experienced raters make decisions while rating the L2 speaking performances. To this end, given that very few studies focusing on speaking performance assessment have been conducted in Turkish and global contexts, this research study sets out to narrow the research gap as regards to the effect of rater experience on score variation and the raters' perception of L2 speaking performances in varying qualities.

1.3. Purpose of the Study

The main goal of this dissertation was to examine the effect of rater experience and L2 speaking performance quality on rater behavior and speaking performance scores. Initially, it aimed to research whether the professional experience of raters (low, medium and high) affects the variability and reliability of rating scores. Secondly, it attempted to uncover the decision-making patterns that raters apply while rating different quality L2 speaking performances. Finally, it intended to find out the sources of score variation that contribute to variability and reliability of speaking performance scores. Utilizing case study mixed-method design, the variability and reliability of given scores were evaluated by quantitative data analysis by including the G-theory approach. Data obtained from rating transcripts and written score explanations were explored through qualitative content analysis.

Considering the quantitative framework, the research questions below were investigated:

1. Are there any significant differences among the analytic scores of low-, medium- and high- quality L2 speaking performances?

2. Are there any significant differences among the analytic scores assigned by low-, medium- and high experienced raters?

3. What are the sources of score variation that contribute most (relatively) to the score variability of the analytic scores of L2 speaking performances?

4. Does the reliability (*e.g.,* dependability coefficients for criterion-referenced score interpretations) of the analytic scores of raters (low, medium and high) differ from each other?

As for qualitative data aspects, the following questions were asked:

5. How do raters make decisions while rating varying quality L2 speaking performances analytically?

6. How does professional experience affect raters' decision-making processes and the aspects of speaking responses they focus on?

1.4. Significance of the Study

As for showing the seeming paradox of raters in a spoken performance assessment, Bachman (1990) clearly states that practically anyone can rate another person's speaking ability, for example, yet; while one rater focuses on pronunciation accuracy, another may find vocabulary to be the most salient feature. This actually shows the extent how much a rater can have an impact on the score variation. Given the complexity of raters' decision-making patterns, empirical research emphasizes the central role of raters in the L2 speaking assessment (Ducasse & Brown, 2009; Knight, 1992; May, 2009; Orr, 2002). Similar to the findings of majority studies as regards to the correlation between rater training and consistent scores, Davis (2016) found that increased rater training contributes to the reliability of

assigned scores. At the same time, the researcher reached the conclusion that the most accurate raters were the ones who gave their decisions utilizing the exemplars rather than the rubrics only. Furthermore, those raters spent more time on the rating process than the least accurate ones. Accordingly, a hasty supposition that accurate raters are supposed to be selected from experienced and native English speaking ones might be misleading.

As can be seen in the relevant literature, there are contradictory findings about raters' professional experience (experienced vs. novice) and the linguistic background (native vs. non-native) (Chalhoub-Deville & Wigglesworth, 2005; Hsieh, 2011; Huang, 2013; Isaacs & Thomson, 2013; Kang, 2008; Kim, 2009a, 2009b; Wei & Llosa, 2015; Xi & Mollaun, 2011; Zhang & Elder, 2011). For instance, Kim (2015) focused on the qualitative analysis of rating decision variations shown from three different rater backgrounds (novice, developing, and experienced). In her study, it was revealed that the developing raters progressed most throughout the training sessions while novice raters did more slowly. Moreover, the experienced raters were the most consistent and reliable in terms of score variation. Similarly, Hubbard et al. (2006) set out to retrieve qualitative data from three expert examiners' decision making patterns while assessing a high-stake speaking test. Their study provided key information about the raters' interpretation of criterion in the rubric as well as the reasons why each examiner opted to select different components.

Considering the linguistic background of raters, Kim (2009b) examined the rating judgments of native and non-native teachers and found internally consistent scores in both groups, which implies that it would be better not to underestimate non-native English speaking raters in terms of overall rating quality when compared to native English speaking ones. The only difference was observed in the commentaries native and non-native English speaking raters gave. For instance, while native English speaking (NES) raters tended to give more intricate and complex opinions about L2 speaking performances, non-native English speaking (NNES) raters relatively chose shallow decision making patterns especially regarding pronunciation and grammatical structures. Similarly, Kim (2009a) found that NES and NNES raters did not differ much as to rater severity but they showed different interaction patterns with test takers. Zhang and Elder (2011) revealed that NES rater features did not overweigh the NNES ones when the score variation issue was considered even if there were

some certain differences between NES and NNES raters in terms of approaching the performance. Furthermore, Wei and Llosa (2015) examined the NES and NNES' rating patterns within the concept of World Englishes. Although there were not any differences in the holistic scores of the speaking test, Indian raters seemed to behave more strictly than American ones in terms of intelligibility of Indian accent. Overall, it would seem that being a NES rater may not always maintain the quality of the rating process. Most importantly, there seems to be increasing recognition of NNES raters in L2 speaking assessment practices. These studies suggest that examining rater's professional experience and background through a combination of facets will probably give researchers a wider perspective on the issue of rater effect on score variation.

A number of studies have been conducted to investigate the facets that might be related to accent familiarity sourced by raters or test-takers. For instance, Huang (2013) examined the issue of accent familiarity and raters' decisions in speaking assessment. The findings of this study indicated that raters having experienced the NNES accent features of the test takers beforehand were found to be less strict and biased while rating the L2 speaking performances. Winke et al. (2012) set out to examine the influence of accent similarity on rating behaviors, one of which showed that the raters tended to be more lenient to the accents that they had contacted either by first language (L1) or L2 experience. Even though the statistical results did not signify a magnitude impact on the score variation, the bias caused by linguistic background would be better taken into consideration especially while planning rater selection. Chalhoub-Deville and Wigglesworth (2005) conducted a study on the English as a Second Language (ESL) test-takers' L2 speaking performances that were evaluated by raters from Australia, Canada, the UK and the US. The findings of this study showed that the score variation was sourced from both rater's linguistic background and task types. As the majority of the previous studies have mostly focused on exploring the score variations stemming from rater's background and accent familiarity, it is of great significance to explore the effect of different rater experience groups (low-, medium-, and high- experienced raters) on the reliability of scores in Turkish higher education context.

The research examining the impact of varying L2 speaking performances on scoring differences in L2 speaking assessment is very limited when compared to the writing

assessment field. In fact, there are not any extensive studies having explored the effect of differing levels of L2 speaking performances on the reliability of assigned scores in both ESL and EFL contexts. As for the distribution of essay quality studies, while most of the studies examined the impact of essay quality on the reliability of scores in ESL context (Brown, 1991; Daly & Dickson-Markman, 1982; Engber, 1995; Freedman, 1981), only few of them investigated this issue in EFL context (Han, 2017; Şahan, 2019). Given that investigating the interaction of various facets matters in terms of understanding the reliability of given scores (Bachman & Palmer, 2010; Weigle, 2002), this study is significant since it aims to bridge the research gap between different speaking performances (low-, medium-, and high-quality responses) and L2 speaking assessment.

Another source of inspiration behind this dissertation was the implications and future directions of a study conducted by Şahan (2018), which essentially focused on two major contexts: a) raters from a single institution, b) raters from various institutions across Türkiye. However, this dissertation aimed to reveal the institutional dynamics of a single EPP as regards to the impact of rater experience and speaking performance quality on score variation and rater behavior. Additionally, this dissertation, and the study carried out by Şahan (2018) collected their main data from the same EPP in Turkish context. Therefore, examining the issues of performance assessment is of wider significance to draw a complete picture of both L2 speaking and writing assessment within the same institution.

The methodological assessment is another point in this study. G-theory, which is built upon classical test theory (CTT), provides researchers with detecting the sources of score variation in a performance based assessment (Brennan, 1992, 2001, 2011). Similar to the methodology of some other relevant studies (Lee, 2005; Xi, 2007; Xi & Mollaun, 2009), this study aimed to utilize a quantitative analysis based on G-theory to investigate rater variation and multiple sources of this variability on assigned scores. Furthermore, this research used verbal protocols to collect qualitative data to be used while observing the evaluation patterns and the rating processes of low-, medium- and high- experienced raters. In fact, the ultimate aim of utilizing verbal protocols was to explore the scoring patterns that raters focused while rating the different quality L2 speaking performances. Since the test takers were speaking during the speaking test, it seemed hardly likely that raters could talk

at the same time. Therefore, the use of verbal protocols enabled the raters to give their retrospective thoughts about their rating behaviors.

Clearly, previous research on the score variability in L2 speaking assessment has mostly examined raters' professional experience, rater training process, and raters' linguistic background especially the accent familiarity. However, there is a very limited amount of research focusing on the score variation sourced from different speaking performance quality. At the same time, the majority of these empirical studies were conducted in ESL context. Therefore, this dissertation is of vital importance in investigating the effect of raters' professional experience on rater decision patterns and rating variability in the context of EFL. Similarly, since the effect of rater experience and L2 speaking performances quality on score variation, and speaking raters' decision making patterns have not been investigated much, this study aimed to fill the aforementioned research gap in L2 speaking assessment research. The findings of this study will provide profound implications for the establishment of a reliable and fair institutional speaking assessment system in higher education contexts.

1.5. Definitions

The key terms that serves the purpose of this study are listed as in the alphabetical order:

Analytic Scoring. In this study, it refers to the detailed assessment of each individual aspect of a spoken performance.

Condition. It refers to the rank of a facet (Shavelson & Webb, 1991).

Criterion-referenced Testing. It refers to a test type assessing students' spoken performance by means of a set of criterion (Brown, 1996).

Decision Study (D-Study). It refers to the optimization of generalizability and dependability coefficient indices to reach the feasible and optimal measurement design (Brennan, 2001).

Facet. A specific aspect in an assessment design (Shavelson & Webb, 1991). It refers to raters and L2 speaking performances in this study.

Generalizability Study (G-Study). It refers to the relative evaluation of various components of the universe score and error variance (Shavelson & Webb, 1991).

Holistic Scoring. It refers to the overall assessment of the student's speaking performance by means of a holistic rubric

Norm-referenced Testing. "A test that measures how the performance of a particular test taker or a group of test takers compares with the performance of another test taker or group of test takers whose scores are given as the norm" (Richards & Schmidt, 2013, p. 363).

Object of Measurement. It refers to the object that exists in the measurement design (Shavelson & Webb, 1991). In this study, it refers to the students.

Rater. It refers to a person who gives scores to either written or spoken performances. In the context of this study, raters are defined as the assessors working as the EFL instructors at a university in Türkiye.

Rater Behavior. It refers to mental procedures through which a rater makes a decision about spoken performances (Davis, 2016).

Rater's In-house Rating Experience. It refers to the number of ratings that a rater has completed in a specific context.

Rater's General Professional Experience. It refers to the number of years spent in teaching and rating spoken performance in the context of this study.

Rating. It refers to the assessment of spoken language performance via rubrics or scales in the context of this study.

Speaking assessment. It refers to scoring an L2 speaking test product that corresponds to a performance. In the context of this study, it is related to L2 speaking assessment in which various approaches and techniques are used for evaluating L2 speaking performance.

Universe. It refers to the mixture of all facets (Shavelson & Webb, 1991).

Variance component. It refers to the degree of a facet in a G-study measurement design (Brennan, 2001).

1.6. Organization of the Dissertation

This dissertation is organized as follows: introduction, literature review, methodology, results, and discussion and conclusion. The second chapter generally reviews theoretical and empirical research on L2 speaking assessment that is relevant to the framework of the study. This chapter touches upon the overall aspects of L2 speaking assessment and fundamental issues in performance tests as well as L2 speaking assessment practices in Turkish higher education context. This section continues with the factors affecting L2 speaking assessment such as speaking tasks, rater's background, rater training, and rating scales. After that, the chapter elaborates into the empirical studies of rater's

experience, text and L2 speaking performance quality in performance assessment, and rater cognition and decision making strategies. Lastly, the literature review chapter provides useful insights into research gaps in L2 speaking assessment. The methodology section first reviews the basics of research design and statistical framework of this study. It then presents information regarding the selection of raters, data collection instruments, data collection procedures, and data preparation. The fourth chapter reports the results and findings of the study by separately organizing as quantitative and qualitative findings. The final chapter mainly summarizes and discusses the results of this study in light of relevant literature as well as the limitations, pedagogical and methodological implications.

CHAPTER II

LITERATURE REVIEW

2.1. Introduction

This chapter consists of six main sections. First, general aspects of L2 speaking assessment and an overview of L2 speaking assessment in higher education in Türkiye are reviewed. Second, factors affecting L2 speaking assessment process are presented, with a particular focus on reviewing speaking tasks, rater's background, rater training, and rating scales. Third, the effect of speaking rater's rating experience on the variability and reliability of assigned scores are reviewed in detail. Fourth, previous research on text and L2 speaking performance quality in L2 performance assessment is examined. In the fifth section, rater cognition and decision making patterns in L2 speaking assessment are discussed. Finally, in the last section, a brief summary of the chapter and a review of research gaps in L2 speaking assessment are provided.

2.2. L2 Speaking Assessment

Direct testing mostly associated with performance assessment is a huge advantage over indirect testing (Weir, 2005). Even if such tests are known to provide more reliable and valid results, test designers or policy makers mostly opt for indirect testing methods for the sake of operational conditions (Fulcher, 2010). Given the operating expenses of performance tests, it is viewed as one of the cost increasing factors. Therefore, indirect tests are still preferred in educational institutions to save time and money while assessing speaking (Fulcher, 2014). Following to the brief comparison of direct and indirect testing, it would be useful to mention that performance assessment has a certain recipe for test designers: describe the target domain, conduct a detailed analysis of the authentic context, create the tasks accordingly, document test-takers' performances and finally make an estimate of student performance using the relevant data (Bachman & Palmer, 2010). In fact, this way of assessment refers to methods through which we can detect and monitor test takers' model performance in authentic domains (Brown, 2004; McNamara, 1996). Specifically,

performance assessment provides optimum conditions and tools for speaking assessment. However, speaking test designers might have to overcome a couple of challenges and need to check whether their tests meet the requirements of typical quality standards in a performance based test (Ross, 2012).

While performance based assessment stands out as a strong way of testing with widely varied tasks and real-life frameworks, it seems to have difficulties in striking the balance between validity and reliability (Kopriva, 2008). This type of assessment brings about some concerns as to generalizing from a model performance to a practical context despite utilizing direct testing tools (Kane, 2012). The particular concern here is clarity and authenticity. In fact, there are two basic issues: a) whether we can really forecast future performances of test-takers and b) whether we can set up a real-life context for testing conditions (Douglas, 2010; Fulcher & Davidson, 2007; Green, 2014). To sustain authentic speaking test items, it is necessary to conduct a detailed analysis of real life contexts. Otherwise, the items would be ill-defined and give unreliable results. Above all, the integration of tasks and constructs in each step can solve two fundamental problems of a performance test: prediction of future performance and authenticity of tasks (Bachman, 1990; Bachman, 2002; Shohamy, 1995).

Given that performance tests have an essential role to assess speaking and writing skills, selecting the right type of test is of critical importance to reliability, validity and fairness of test results. In fact, determining the direction and purpose of a test is the most critical stage of test development. The characteristics of language tests can be categorized as ‘test use’, ‘content’ based on proficiency and syllabus, ‘norm and criterion referenced interpretation’, and ‘testing method’ (Bachman, 1990). Each test is supposed to be designed to perform a specific function, from placing test takers in a language level to assessing overall language skills (Hughes, 2003). As such, it is useful to be aware of the functions of assessment and test types. For instance, when teachers aim to evaluate their students’ ongoing learning experience, they utilize formative assessment, namely, informal assessment. However, summative assessments are used to measure students’ overall or end course performance (Brown, 2004; Fulcher & Davidson, 2007). Even a simple low stakes test can change students’ life negatively by placing them in the wrong classroom level. Thus,

the probability that high stakes tests affect students' lives completely makes summative assessment superior to formative assessment (Douglas, 2010).

The main characteristic of speaking assessment tasks is based on a causal relationship between four aspects: 'performance', 'the test system', 'the test taker' and 'the scoring system'. In fact, it is unlikely that a speaking test will not be affected by factors such as test administration, rater behavior, task type, test format and scoring rubric (O'Sullivan, 2013). In addition to these aspects, spoken interaction primarily needs to be considered in every stage of L2 speaking assessment design as the nature of speaking is immensely complicated and depends on numerous variables such as body language, turn-taking and recasting (Galaczi, & Taylor, 2018). The fact that performance assessment tasks have complex characteristics signifies the importance of building clear rating procedures together with professional development support for raters before and after each test (Douglas, 2010). Therefore, test builders need to adopt a comprehensive framework in which they have to manage task characteristics such as setting, time, rubrics, format and input. For instance, whether sufficient time between the tasks is provided to test takers might affect the quality of the response, directly resulting in score variation (Bachman & Palmer, 2010).

2.2.1. Reliability and Validity

In assessment, reliability and validity are two crucial elements that test designers have to strike the balance to minimize the variation risks. This actually highlights the issue of reliability and validity in speaking assessment in which test designers gain knowledge of test takers' L2 speaking skills through performance tasks (Hughes, 2003). Given that raters are the potential source of variability in speaking assessment, subjectivity will be unavoidable unless test designers do not take sufficient measures to ensure reliability and validity. While rater training and standard cut scores can be given as examples of measures for building reliability, test purpose, task specifications, rating criteria and washback effect are the components of validity steps (Brown & Hudson, 2012; Luoma, 2004; McNamara, 1996). As for the concept of reliability, it has an essential principle: as long as a test is administered to the same group of test takers by setting up more or less similar conditions

each time, and resulting in consistent, dependable and fair scores, a test designer can attribute this as the reliability of scores (Bachman, 1990; Fulcher & Davidson, 2007; Hughes, 2003). While the true score refers to the expected score that a test taker would present as the true ability, the measurement of error is concerned with the score that is not related to the test taker's ability being tested but other factors (Brown & Hudson, 2012; Luoma, 2004). Therefore, according to this definition, it would not be wrong to claim that more errors in a measurement might refer to less reliability of given scores (Henning, 1987).

Given the place of human related subjectivity in language assessment, the issue of rater reliability is present with inter-rater and intra-rater reliability (Brown, 2004). The fact that two raters yield consistent scores signifies inter-rater reliability. However, as for intra-rater reliability, the focus is on one rater' scores with the same set of test takers' performances but on a different period of time. Basically, while the former is interested in the consistency of two raters' scores, the latter tries to reveal whether there are inconsistencies in one single raters' scores (Bachman, 1990; Brown, 1996; Fulcher, 2014). Considering the complexity of L2 speaking assessment, ensuring excellence in inter-rater reliability and intra-rater reliability entails some risks and thus requires a few procedures. As such, Luoma (2004) suggests an analytical process with three steps that provides reliability in speaking assessment: a) raters' internal consistency, namely, intra-rater reliability stage where test designers investigate whether raters show the same level of agreement in a period of time, b) inter-rater reliability stage where two raters agree on their assigned scores by using a detailed and clear set of rating criteria, and c) the third one is parallel form reliability stage where testers need to analyze and compare the scores that raters assigned by means of cross-tabulation. In fact, as described by Luoma (2004), once test designers implement these three quality stages, a dynamic and ongoing reliability check system will be constructed while carrying out speaking assessment.

Having reviewed the practical measures for sustaining reliability in the previous paragraphs, I can speculate that the quality cycle of a speaking assessment process needs to be complemented by reliability and validity (Fulcher, 2014). To understand the concept of validity in any language assessment types, two key points have to be at the top of the to-do-list: a) determining the purpose of the test and b) considering the appropriateness and

practicality of the test (Luoma, 2004). While reliability investigates to what extent the score variation stems from measurement error and other reasons, validity aims to determine to what extent a test is appropriate, useful and appropriate (Bachman, 1990; Brown, 2004; Fulcher & Davidson, 2007). Unlike reliability, validity is engaged in the correlation between particular language skills and test performance (Bachman, 1990). Similar to the previously mentioned definitions, McNamara (1996) highlights two key points regarding validity: "how and how well we can generalize from the test performance to the criterion behavior" (p. 16). Adding up all these definitions, testers should not underestimate the power of utilizing both reliability and validity quality stages at the same time while designing L2 performance assessment tests (Bachman, 1990).

Another important but complex issue in L2 performance assessment is fairness. While the "procedural fairness" prioritizes equal opportunities and similar conditions for all test takers, the "substantive fairness" is concerned with the bigger picture of the testing system in question and the relationship between the system and other various stakeholders (Kane, 2010, p. 178). These two definitions are specifically interrelated to validity. The complexity of fairness issues especially becomes salient in performance assessment as regards to the discussion of equality or equity. Given the operational difficulties of conducting a performance based test, treating each test taker as one person would give fair results with more equity. However, this would also pose some risks in terms of comparing the results. In fact, the notion of fairness in language assessment is still disputable since there is a close similarity between validity and fairness (Davies, 2010).

2.2.2. An Overview of L2 Speaking Assessment in Higher Education System in Türkiye

A two-stage test system, organized by the Student Selection and Placement Center places students into the departments of universities. These tests are high-stakes and based on multiple-choice questions. First of all, students take the first test called the Basic Proficiency Test in which they answer basic course subjects such mathematics, social sciences, science and Turkish. The students getting an acceptable score from this test can enroll for two-year

degree programs. In the second stage, students sit for the Field Qualification Test in which they answer questions based on their field groups. To enroll for four-year degree or bachelor programs, the cumulative scores of these tests, and the average score of a high-school diploma are calculated and after that students make a list of departments they want to study in a rank order and upload their preferences to the automation system. Finally, this center automatically places students according to their rank order (Kitchen et al., 2019).

As for the medium of instruction, YÖK regulates whether Turkish, English or mixed Turkish-English medium is used. However, the proportion of language mediums either 100% or 30% in English is determined according to the regulations of departments and universities since faculty members need to meet the minimum requirements for lecturing in English. In addition to this, with the advance of a recent regulation, the access to EPP has been limited to only the students registering for full (100%) and partial (30%) English Medium Instruction (EMI) departments. However, students of Turkish medium of instruction can voluntarily enroll in an EPP based on the board decision of this academic department. As a result of this, universities have been trying to increase the capacity of EPPs and the number of either full or partial EMI programs (British Council, 2015).

EPPs in Türkiye mostly offer a mixture of English for general and academic purposes based on Common European Framework of Reference (CEFR) levels. Given the assessment system, it can be said that most of the EPPs prepare their own low-stakes and high-stakes exams since institutional needs and situations may vary. That is to say, there are different assessment and evaluation protocols in EPPs across the country. However, the basic structure of the student acceptance and placement system in EPPs is more or less similar in each school: English proficiency and placement tests. The students who get the acceptable score from the proficiency test have access to departments while the ones that do not receive the passing grade sit for the placement test and start their EPP. Therefore, EPPs have been exposed to widespread criticisms mainly due to this gatekeeper role between the students and academic departments. These criticisms are mostly about curriculum design, the efficacy of instructors and assessment quality. Students receive intensive four skills (reading, listening, writing, and speaking) instruction with emphasis on the communicative aspects of English language. However, less integration among four skills stands out as a serious threat

to the curriculum and assessment quality. In addition, the problems stemming from non-standardized tests occur since EPPs mostly lack the quality steps for validity and reliability and at the same time face disadvantages such as crowded class population, time constraints and teacher burnout. Compared to reading, writing, and listening skills, the paradox of less attention on the assessment of speaking skills seems to be one of the most challenging L2 speaking assessment issues in EPPs (British Council, 2015). Given the efforts of YÖK to foster internationalization and the presence of interaction in EMI classes, the instruction and assessment of L2 speaking is of critical importance to the quality of higher education programs in the country.

L2 speaking assessment necessitates a methodical route to reliability and validity due to the complexity of speaking skills constructs, the interaction between human raters and test takers, and operational concerns in testing (Fulcher, 2014). In Türkiye, most of the speaking test items are prepared either by the testing units or examination workgroups within EPPs. However, these in-house procedures are generally far from detailed and well-planned speaking assessment principles. For instance, some of the EPPs follow a double scoring policy with clear rating rubrics while others may just adopt one rater policy and even borrow speaking rubrics or test items from course book publishers and other institutions. Another side of the issue is that some of the EPPs might not have a proper testing unit, which makes the scenario even worse because this means that there are not any protocols regarding the management of the EPP tests. According to the findings of the British Council report on the state of higher education in Türkiye, only reading and writing skills tests were assessed in English proficiency tests; yet, listening and speaking skills were neglected in the assessment table. Furthermore, instructors contributing to the data collection of the British Council report stated that there were some infringements relating to adherence to academic integrity standards in EPPs since some of the instructors reported that they were forced to change or inflate some unsuccessful students' writing or speaking scores. In fact, this would ruin the validity and the reliability of these tests (British Council, 2015). Therefore, it is clear that there is no unity of quality assurance as regards to L2 speaking assessment in EPPs across the country. There is no doubt that these non-standard procedures will cause reliability issues in speaking assessment, resulting in a domino effect on the other skills in the whole assessment table.

2.3. Factors Affecting L2 Speaking Assessment

Assigned scores are dependent on varying components that interact with each other. The scoring variation in speaking assessment can be attributed to factors such as test takers, interlocutors, tasks, rubrics, raters, and the quality of performance itself (McNamara, 1997). More specifically, Bachman (1990) explains the effect of factors on performance test scores by means of three aspects: 'test method facets', 'personal attributes', and 'random factors'. While the first factor is classified as 'the testing environment', 'the test rubric', 'the input', 'the relationship between input and response', - the second one refers to test takers' background such as race, gender, and subject knowledge. The common point between test method facets and personal attributes is that they both might have 'systematic effect' on given test scores. However, the third factor is 'unsystematic' and may arise if sudden changes or situations happen. Given the possible effects of various factors on the speaking assessment process, examining the relationship across them will contribute to ensuring the reliability of speaking test scores (Luoma, 2004). Due to raters' fundamental characteristics, it would not be wrong to express that raters are one of the most dominant factors that are responsible for scoring interpretation and variation (Bachman, 2004; Fulcher, 2014).

This section of the chapter elaborates into details of the factors affecting the variability of speaking test scores. The topics are as follows: tasks in speaking assessment, rater's background, rater training, and rating scales. Following this, the section gives a detailed review of the relevant literature of rater's rating experience and text and L2 speaking performance quality in L2 performance assessment. Furthermore, relevant studies of rater cognition and decision making patterns that raters employ are reviewed as the final factor affecting L2 performance assessment. In some of the sections, empirical studies from L2 writing assessment were presented as some issues were relatively understudied in second speaking assessment field. Showing the overall picture, Table 1 below summarizes the reviewed studies in this chapter.

Table 1

A general overview of reviewed studies

The scope of studies	Number of studies	Relevant studies
Tasks in L2 Speaking Assessment	11	Elder et al., 2002; Fulcher & Reiter, 2003; Huang et al., 2018; Teng, 2007; Iwashita et al., 2001; Kim & Craig, 2012; Khabbazbashi, 2016; Lumley & O'Sullivan, 2005; Ockey et al., 2019; Tavakoli, 2009; Weir & Wu, 2006
Rater's Background	13	Caban, 2003; Carey et al., 2011; Douglas & Myers, 2000; Huang, 2013; Huang & Jun, 2015; Huang et al., 2016; Kang et al., 2019; Kim, 2009a; 2009b; Lumley, 1998; Wei & Llosa, 2015; Winke et al., 2012; Zhang & Elder, 2011
Rater Training	7	Kang et al., 2019; Lumley & McNamara, 1995; Papajohn, 2002; Stitt et al., 2003; Wigglesworth, 1993; Xi & Mollaun, 2011; Yan, 2014
Rating Scales	6	Brown, 2007; Brown, 2006; Chuang, 2009; Fulcher et al., 2010; Isaacs & Thomson, 2013; Upshur & Turner, 1995
Rater's Rating Experience	11	Barkaoui, 2010a; Bonk & Ockey, 2003; Cumming, 1990; Davis, 2016; Delaruelle, 1997; Kim, 2015; Leckie & Baird, 2011; Lim, 2011; Sakyi, 2003; Song & Caruso, 1996; Wolfe et al., 1998
L2 speaking performance and Text Quality	6	Brown, 1991; Daly & Dickson-Markman, 1982; Engber, 1995; Freedman, 1981; Han, 2017; Şahan, 2019
Rater Cognition and Decision Making Patterns	8	Ang-Aw & Goh, 2011; Brown et al., 2005; Cai, 2015; Chalhoub-Deville, 1995; Gebril & Plakans, 2014; Gui, 2012; Orr, 2002; Pollitt & Murray, 1996
Total	**62**	

2.3.1. Tasks in L2 Speaking Assessment

Given the main character of speaking skill, a mechanism with its own system and principles, namely speaking tasks, is necessary to assess the performance of test takers. Otherwise, so-called speaking assessment would turn into the format of saying something or just talking sessions. Thus, test designers need to follow a strict procedure while creating speaking tasks to reveal testers' optimum performance (Luoma, 2004). Initially, the focus of speaking assessment task design had been on two points: restricting the type of speaking exam tasks and evaluating the positive and negative sides of these tasks (Madsen, 1983). However, this approach to speaking assessment task design has evolved into the inclusion of using novel assessment models (Bachman, 1990; Bachman & Palmer, 1996; Nunan, 1989). The starting point of Bachman and Palmer's model was to describe the language use by classifying it as 'real-life domains' and 'language-instruction domain'. In fact, the presence of language use was salient in terms of building a bridge between the test item and real life language itself. Compared to earlier models solely providing ready to use speaking tasks, this framework model was on the stage to create a tool set describing the characteristics of aforementioned domains. Therefore, test designers can use this model as a handy checklist to adapt various test situations (Bachman & Palmer, 1996).

Task difficulty is another critical issue in L2 speaking assessment since it might directly affect the quality of test takers' performance. Thinking outside the box of assessment, our concerns generally come from the situations or contexts that we have never experienced or seen before. Similarly, the same concern that we can also call 'communicative stress' in speaking assessment is quite related to three aspects: a) context familiarity, b) test takers' background knowledge, and c) task type. L2 speaking test items in English prepared without taking these aspects into consideration could cause stress even for a native speaker of English. Given the pressure stemming from the process of L2 learning, it is more likely that this type of task would be relatively more stressful for a non-native speaker of English test taker. As such, task difficulty issues such as complexity, familiarity and background knowledge could be addressed while designing speaking tasks (Brown & Yule, 1983; Skehan, 1998). Although test designers need to be aware of such items regarding task difficulty, the sources of task difficulty might not be very clear since

there are various interrelated factors such as raters, test conditions and test takers in the nature of speaking assessment (Bachman, 2002; Elder et al., 2002; Luoma, 2004). Task type is another issue in L2 speaking assessment. Having mentioned before, the stage where task types are evaluated within the framework of models is crucial in the design of speaking tasks. Thus, task types and these frameworks are interrelated. Given the relative popularity of interviews across speaking task types, they seem to be opted more frequently than other various task types such as picture prompts, presentations and short stories because of operational factors (Fulcher, 2014). In that context, numerous studies examined the effect of task difficulty on the variability of speaking test scores (Elder et al., 2002; Fulcher & Reiter, 2003; Iwashita et al., 2001; Tavakoli, 2009; Weir & Wu, 2006), the impact of task topics on assigned scores (Huang et al., 2018; Khabbazbashi, 2016; Lumley & O'Sullivan, 2005), and comparison of various task types (Teng, 2007; Kim & Craig, 2012; Ockey et al., 2019).

The possible impact of task difficulty on test takers' performance has been disputable among scholars. With this in mind, it would not be wrong to state that there are two poles in this matter. Iwashita et al. (2001) investigated whether the aspects of tasks and performance conditions affect task difficulty. The task dimensions and performance conditions were adapted from the framework developed by Skehan (1998). A number of 193 ESL students from EPP courses in Australia participated in the study. The narrative tasks in this study were designed as less difficult and more difficult dimensions. For instance, while the students were given 3.5 minutes for reading the instructions and preparing the task in one round, they were given solely 0.5 minutes for both of them in another round. In contrast to the anticipated findings of Skehan's framework, this study revealed no variation between the task dimensions and task difficulty. However, the researchers signified the importance of researching the concept of task difficulty by utilizing different task types in testing situations. In the same vein, Elder et al. (2002) explored whether there is a relationship between task complexity and task difficulty as well as the opinions of test takers on task difficulty. A total number of 210 ESL students contributed to this study by taking the in-house version of the TOEFL test and the speaking test. As for the oral test, the participants responded to eight narrative tasks: four less demanding and four more challenging tasks according to the model proposed by Skehan (1998). In addition, the test takers filled out a questionnaire eliciting about perceptions of and attitudes towards task difficulty. The data were used to corroborate

the findings from the quantitative data. Fourteen raters, having received training before, scored the test takers' performances. As for the relationship between task difficulty and complexity, this study revealed striking contrasts to the model proposed by Skehan (1998), which actually claimed that as the complexity of the task increases, the difficulty level increases. Indeed, similar to the findings of the previous study, there were no differences in task difficulty across task dimensions. The qualitative findings based on the test takers' perceptions also showed no differences in terms of task difficulty. Despite these findings, the researchers opted to be skeptical about this difficulty issue since the type of task, rater effect and test takers' subjectivity might have an effect on task difficulty.

However, some of the studies focusing on task difficulty contradict the finding that there is no relationship between task difficulty and assigned scores. In this sense, Fulcher and Reiter (2003) carried out a study to uncover the possible differences of task difficulty as to task conditions and L1 cultural aspects. A total number of six speaking tasks with varying levels and difficulty were used in this study. As the nature of study is based on pragmatic and cultural effects, 32 students with L1 English speaking background and 23 L1 Spanish speaking students were the participants. The results of the statistical tests indicated that there is a systematic relationship among task dimensions, language use, and pragmatic aspects. Namely, pragmatic and cultural aspects might have an effect on the difficulty of tasks. In another study, Weir and Wu (2006) examined the parallel forms reliability and content validity of an intermediate level speaking exam consisting of three tasks: reading aloud, responding to questions and picture prompt tasks. A number of 120 EFL students from Taiwan responded to three types of tasks in three trial versions of this test. As for data analysis, while ANOVA, factor analysis and multi-faceted Rasch measurement were utilized for the quantitative data, raters' opinions on the difficulty of tasks were used for the qualitative data. Although the quantitative findings revealed a parallelism across the three versions of the test, raters reported that test Version 1 showed differences as regards to task difficulty. Therefore, the researchers highlighted the importance of adding qualitative aspects in the research design of task difficulty studies. Utilizing retrospective data collection techniques, Tavakoli (2009) investigated the issue of task difficulty from the perspectives of test takers and instructors. Ten intermediate level students learning English at a college and 10 instructors working at this institution participated in this research. The participants

evaluated the tasks in terms of cognitive, linguistic, clarity, information load, structural, and affective dimensions. Unlike the relevant studies examining task difficulty, this study was unique in terms of comparing the test takers' and teachers' perceptions of task difficulty. More importantly, the research provided insights into the underlying reasons of task difficulty. As for the findings, both participating groups commented on the perceived task difficulties. Given the reasons for these difficulties, one of the participants, for instance, reported the presence of unknown words and structures. All in all, even though there were some points where test takers and instructors did not agree on the task dimensions, the fact that the test takers' and instructors' opinions mostly correspond with each other could be an example model for further research.

Task topic is one of the most crucial aspects in L2 speaking assessment. Using various statistical analyses, Lumley and O'Sullivan (2005) examined whether task topics that are familiar with female and male cause variations in the scores. A total number of 894 students taking an exit level high-stakes English test in Hong Kong after the university graduation contributed to this study. This speaking test was tape-recorded and the gender of the interlocutors were manipulated for the study. In addition, task dimensions were adapted as female-based and male based topics such as housing, horse racing, and leisure. As for the examiners, 30 expert raters awarded scores to test takers' performances. This study revealed that there was not a considerable variation in test takers' scores due to the gender of the interlocutor. However, task topics with gender bias tended to have more effect on the score variation than other factors. The researchers underlined the importance of designing contextual aspects in speaking tasks because the selected topics, the gender of audience and interlocutor might cause variations in test takers' performances. Utilizing both quantitative and qualitative data collection techniques, Khabbazbashi (2016) investigated whether test takers' prior knowledge about topics and task topics have effects on the score variation in the speaking part of IELTS. A total number of 81 test takers speaking Farsi as a mother tongue participated in the study. According to the findings of this comprehensive study, test-takers' prior knowledge of the topic did not show a significant result because it was reported by raters that follow-up questions assisted the test takers when they did not have any information regarding the question. Another point that the raters made was that when the test takers were faced with an unknown topic, their performance quality decreased in terms of

grammar and vocabulary range. Thus, topic selection and test takers' prior knowledge of the task topics need to be approached carefully and controlled to achieve reliability and fairness. In another study, Huang et al. (2018) conducted a study to explore the relationship between background topic knowledge and test takers' performance in a high stakes speaking test. 352 EFL learner test takers from varying degrees and age groups contributed to the study. Utilizing path analyses, the researchers found out that there was a strong relationship between topical knowledge and speaking test performances. According to the findings, topical knowledge showed varying effects on the integrated and independent tasks. The findings of this study provided valuable implications for the use of pre-task input in integrated speaking assessment tasks.

The effect of various task types on the performance of test takers has become one of the focal centers of interest in L2 speaking assessment (Hirai & Koizumi, 2009; Iwashita, 1998; May, 2011; Norton, 2005; Swain, 2001; Van Lier, 1989). Teng (2007) explored whether EFL learners' scores showed variation in terms of accuracy, fluency and complexity across three types of speaking test tasks, which were presentation, picture description and answering questions. 30 EFL learners with high-intermediate level participated in the study. Using a holistic speaking test scale, two trained raters assigned scores to the test takers' tape-recorded performances. On top of that, relying on the frequency of types, mistakes and syllables in the clauses, the raters analyzed the transcription of the recorded performances to reveal the differences of accuracy, fluency and complexity. The study also investigated the test takers' attitudes towards the three tasks by means of an affective dimension questionnaire. Based on the analysis of one-way ANOVA, it was found that the test takers' holistic score did not show any differences across three task types. Nevertheless, there were significant differences as regards to the complexity and fluency of the responses. For instance, test takers produced more complex and fluent sentences in the task of answering questions since this task was direct, structured, and more demanding compared to the presentation and picture description tasks. As for the affective dimensions, the test takers reported that they felt more mental pressure in answering questions than the other two tasks. At the same time, the test takers favored the picture description task since they gave opportunities of visual support. Utilizing experimental research design, Kim and Craig (2012) compared the webcam based live speaking test and face-to-face speaking test in terms

of reliability and validity aspects. 40 test takers took both web-based and face-to-face speaking tests in a different time period. For this interview design, randomly selected groups with 20 test takers each were created in this time period. As for the qualitative data, 10 of them volunteered for the focus group interview. Two trained raters assigned scores following each session. The results revealed that there was no significant difference of test takers' performances in both task types. Supporting the quantitative findings, the majority of test takers underlined the similarities between webcam based and face-to-face speaking tests. Namely, this new type of speaking test task type was as favorable as the old classical one, face-to-face interview. Similar to this study, utilizing mixed methodology research design, Ockey et al. (2019) explored the feasibility of a Skype based speaking test in the US and China. As for the participants and types of tasks, 72 test takers from the US and 74 test takers from China completed four tasks: group discussion, replying to short questions, summarizing, and presentation. According to the overall findings, both participating groups were satisfied with the Skype based speaking test. However, the problems stemming from technical aspects, especially in China, disrupted the functionality of the test. During the focus group sessions, some of the participants mentioned that video-based speaking assessment could decrease the stress of communication with an examiner or interlocutor. This finding is salient in terms of affective dimensions. Therefore, it could be mentioned that the use of technology in speaking tasks is becoming increasingly promising and feasible as regards to reliability and validity although technical limitations might pose some risks to test management.

As can be seen in the review of aforementioned studies, the speaking task is comprehensive and needs to be approached in detail as it is interrelated to other aspects such as raters, test takers, culture, gender, and technology. While some of the studies revealed non-significant results of task difficulty and score variation (Elder et al., 2002; Iwashita et al., 2001), some of them highlighted variation stemming from task difficulty (Fulcher & Reiter, 2003; Tavakoli, 2009; Weir & Wu, 2006). In fact, these studies signify the necessity of conducting more research into task difficulty and score variation. In addition, task topics in speaking assessment are potential sources of having an impact on test takers' performance. Gender biased topic selection (Lumley & O'Sullivan, 2005), the effect of background knowledge and (un)familiarity with the topic (Huang et al., 2018; Khabbazbashi, 2016) need

to be considered while designing speaking tasks. Task type also affects the quality and quantity of test takers' performance. For instance, test takers can give more fluent and complex answers in the interview task than the picture description or presentation tasks (Teng, 2007). In the institution where the main data of this dissertation were gathered, speaking topics are primarily determined based on the topics covered in the speaking section of the main course book. Then, these topics are evaluated by an independent committee in terms of task difficulty. Then, the test takers are provided mock exams before the actual final speaking exam to make them familiar with the tasks and topics. Thus, the researcher collected the main data from this institutional speaking final exam to eliminate risks and disadvantages stemming from task difficulty, task topic and type.

2.3.2. Rater's background

The research on whether a certain rater background (different occupation groups or native and non-native speakers of a language) award more or less consistent and severe scores might have valuable implications for score variation and rating patterns in language assessment (Brown, 1995; Duijm et al., 2017). For instance, Lumley (1998) examined whether ESL raters assigned scores more leniently than doctors in Occupational English Test, which is an ESP test designed for medical employees in Australia. A total number of 10 ESL raters and 9 doctors awarded scores to 20 test takers' recorded speaking performances. The findings of the study revealed that both rater groups showed similarities in terms of scoring. However, there were some tendencies of assigning higher or lower scoring patterns by individual raters in both groups. Given the findings of sub categories of the test, ESL raters tended to give lower scores than the doctor raters. In another study, utilizing ethno-methodological techniques, Douglas and Myers (2000) also explored the possible differences between language expert raters and non-language profession raters. As for the subjects, veterinary professionals and applied linguists evaluated the video recording of veterinary students who were interacting with their clients. Both rater groups separately gathered and discussed the test takers' recordings in detail. The results showed that veterinary professionals focused on the aspects related to veterinary subjects whereas language professionals concentrated on language related aspects.

Douglas and Myers (2000) and Lumley (1998) investigated the points of contention between language instructors and other professionals as regards to rating behaviors. Similarly, but with more rater groups, Caban (2003) explored the possible differences and score variations across four different rater groups: a) ESL MA students (L1 English speakers and L1 Japanese speakers), b) L1 English speakers who are teachers in a secondary school, and c) Peer ESL students with L1 Japanese background. Four students' performances were recorded and delivered to the raters via a CD. 83 raters evaluated four students' performances using a Likert-scale with 15 options ordering from very poor to excellent. The main categories in this scale were Grammar, Fluency, Pronunciation, Content, Compensation, Language, and Intelligibility. The results retrieved from the scale showed similarities in all rater groups. For instance, while all raters awarded low scores to the fluency category, they assigned balanced scores to the other categories. Looking into details of the rater differences, the researcher used the FACETS, which is a program used in multifaceted Rasch measurements, showing that L1 English speaker MA students and L1 English speaker school teachers gave higher scores to pronunciation. However, L1 Japanese speaker MA students awarded lower scores to grammar and pronunciation. The peer ESL students gave relatively higher scores to the fluency and grammar categories. The most striking finding of this study was to signify the gap between the scale and FACETS analysis. Therefore, test designers are to consider the potential differences among different rater groups in speaking assessment.

Caban (2003) specifically focused on the raters' professional background and disclosed particular differences among rater groups. Given the comprehensive findings, the broad spectrum of subjects from ESL professionals to peer students was one of strengths of this study. Similar to Caban's study, Huang (2013) focused on two main areas: the impact of raters' Chinese accent familiarity and raters' professional background on rating behavior. There were three major rater groups: a) Non-language instructors having no contact with Chinese accent before, b) Non-language instructors having direct contact with Chinese accent, and c) Language instructors having direct contact with Chinese accent. In this study, while language instructors refer to teachers with EFL and ESL experience and background, non-language instructors are either college students or personnel working at universities. There were not any significant effects of raters' accent familiarity and professional background on rating behavior. However, the raters' comments on the issue provided

contrasting findings. Some of the raters stated that they awarded higher scores due to their familiarity with the accent and thanks to their EFL/ESL professional background.

Another popular research interest of L2 speaking assessment research is the issue of accent familiarity. In this vein, Carey et al. (2011) investigated whether the familiarity of interlanguage pronunciation patterns by different rater backgrounds has an impact on the score variation in a high stake speaking test called Oral Proficiency Interview. A total number of 99 IELTS OPI raters were formed from five different test centers: India, Hong Kong, Australia, New Zealand, and Korea. Sixty speaking performances of OPI were collected from the test centers located in Korea, Hong Kong, and India. According to the findings based on pronunciation scores, the raters' accent familiarity level had an impact on the scores, resulting in higher scores and lower scores in the cases of unfamiliarity. Similar tendencies were observed in the correlation between the test location and the assigned pronunciation scores.

Huang and Jun (2015) carried out a study investigating whether the raters' varying L1 background and language experience had an impact on their ratings of foreign accent in terms of severity, leniency, and reliability. As for the subjects, 64 Mandarin Chinese speakers from different L2 backgrounds contributed to the study. Their arrival period to the US as immigrants determined their groups: ages 5-11, ages 12-16, and ages 17-25. However, only one group (n=14) was the NES control group. Three rater groups were formed based on a set of criteria in line with the research questions: a) advanced NNES raters, b) low-experienced NES raters, and c) high-experienced NES raters. While deciding on the level of NES raters' experience, the researcher required the raters to report their level of foreign accent familiarity. According to the analysis of interrater reliability, both high-experienced and low-experienced NES raters were more reliable than the advanced NNES. Low-experienced NES raters awarded lower scores than the high-experienced NES and advanced NNES raters. As for the ability of differentiating NES and NNES speech samples, high-experienced NES raters performed more successfully than the advanced NNES and inexperienced NES raters.

Huang et al. (2016) conducted a study to explore the impact of accent familiarity by comparing the behaviors of raters from different groups. In total, 48 raters contributed to this study from three separate rater groups: Spanish-Heritage, Spanish-Non-Heritage, and Chinese-Heritage. The concept of heritage here refers to raters' family, cultural and linguistic ties. Utilizing a Likert-scale ranging from 1 point (poor) to 7 points (native English speaker level), the raters assessed 28 test takers' performances taken from TOEFL iBT public version. Based on the quasi-experimental research design, the study revealed that there were differences in the perception of accent familiarity between Spanish and Chinese rater backgrounds. To illustrate, Spanish-Heritage and Non-Heritage groups detected Spanish accent patterns more effectively than the Chinese-Heritage group. However, when examining the overall numerical ratings in detail, raters from both heritage groups showed differences from the non-heritage group in terms of feeling closer to foreign accents. According to qualitative findings of this study, raters disclosed that they tended to give higher scores when they came across a familiar accent.

Kang et al. (2019) investigated the relationship between score variance and rater background as well as the effect of basic rater training. A total number of 82 novice raters evaluated speaking performances retrieved from TOEFL iBT speaking tasks. There were significant differences between NES raters and NNES raters as to the severity of assigned scores. While NES raters were more tolerant, NNES raters were harsher. Raters' familiarity with non-native accents was another focus of this study. It was revealed that the raters having stronger links with non-native varieties of English were less severe while rating speaking performances. At the same time, this study examined the impact of raters' stereotypes of NNES varieties, resulting in less reliable and valid scores.

However, the comparison of NES and NNES did not always cause score variations, yet different rating patterns. Using quantitative and qualitative research methods, Kim (2009b) aimed to research native and non-native speakers of English raters' rating patterns while assessing a sample of non-native English speaker Korean students' speaking test performance. A specific speaking test with eight tasks was designed for this study. In addition to this, a rating scale ranging from 1 (unsuccessful) to 4 (almost always successful) was developed. As for the subjects, 10 Korean students as test-takers, 12 NES raters from

Canada and 12 NNES raters from Korea contributed to the study. The quantitative analyses showed that there were not any significant differences in the both rater groups' rating patterns. However, the qualitative findings revealed that NES raters produced more complex notes than NNES counterparts while filling out the assessment criteria. In the same vein, to investigate the differences between NES and NNES teachers' rating behaviors and patterns, Zhang and Elder (2011) designed a mixed-method study with 19 NES and 20 NNES raters and speaking performances retrieved from 30 test takers, who took a high-stake speaking test organized in China. The raters in this study were English instructors working at different universities in China and Australia. A holistic rating scale ranging from 1 (very poor) to 5 (excellent) was utilized. Quantitative results showed that there were not any significant differences of score variation between NES and NNES raters. Nonetheless, qualitative results disclosed certain dissimilarities across both rater groups. For instance, NES gave more elaborate explanations about the test takers' performance and focused on the communicative aspects and task strategies. However, NNES raters mostly concentrated on grammatical forms of the speech samples. Given the situation of education context in China, the researchers attributed these differences to social, cultural, and educational factors.

Both Kim (2009b) and Zhang and Elder (2011) revealed that there were not any significant differences between NES and NNES rater groups in terms of assigned scores. Nevertheless, qualitative findings disclosed some differences as regards to amount and frequency in raters' reports. Similar to these studies' findings, Wei and Llosa (2015) carried out a study whether the differences between American and Indian raters have an impact on the score variation and rating behaviors. The study was based on a mixed-method research design utilizing both quantitative and qualitative methods. The L2 speaking performance data were gathered from 240 test takers' TOEFL iBT speaking test performances. As for rater subjects, three American and Indian raters from each participated in the study. According to the quantitative findings, score variation was not observed between the two rater groups. Examining verbal report analysis, the researchers revealed some differences between American and Indian raters' rating processes. While American raters found Indian accent speech samples challenging, Indian raters were relatively more successful at detecting Indian accent patterns. The most salient aspect of this study was to focus on not only non-native accents but also lexical, grammatical, discourse, and cultural aspects. In fact, Indian

raters were better at comprehending these aspects than the American fellows. Therefore, this study gave valuable implications for rater training, as well.

The aforementioned studies focused on the impact of raters' L1 background on the rating patterns and score variation. However, Winke et al. (2012) examined the effect of raters' L2 background on score variation while evaluating test takers' L2 speaking performances obtained from TOEFL iBT speaking test. While raters were L2 speakers of Spanish, Chinese, and Korean, test takers were L1 speakers of Spanish, Chinese, and Korean. A total number of 107 raters and 72 test takers contributed to the study. According to the findings of multi-faceted Rasch measurement, L2 Spanish and Chinese raters gave higher scores to L1 speakers of Spanish and Chinese, respectively. However, there were not any significant differences for the case of L2 Korean raters and L1 Korean test takers. Unlike other studies researching raters' L1 background, this study explored the effect of raters' L2 background on rating behaviors. As such, the researchers underlined the importance of researching raters' educational and cultural experiences about the language in question.

To sum up, it can be seen from the reviewed studies that the majority of the studies focused on the impact of raters' L1 background on score variation and rating patterns, especially the differences between NES and NNES raters (Carey et al., 2011; Huang & Jun, 2015; Huang et al., 2016; Kang et al., 2019; Kim, 2009b; Wei & Llosa, 2015; Zhang & Elder, 2011). In addition to this, the relationship between raters' professional background and rating behaviors was investigated (Caban, 2003; Douglas & Myers, 2000; Huang, 2013; Lumley, 1998). One of them investigated the impact of raters' L2 background on the rating process (Winke et al., 2012). Even though the raters in this dissertation had varying degrees of education ranging from BA to PhD, they reported to have major degrees from ELT and ELL departments. Furthermore, they were all employed as EFL instructors at the university where the l2 speaking performance data were gathered. All the test takers' L1 were Turkish. Therefore, NES instructors and foreign students were not included in this study to minimize any possible effects stemming from participants' background.

2.3.3. Rater Training

Given the basic nature of human beings, ratings can be based on subjective judgments rather than a rating scale unless precautions are taken (Brown, 2004). Performance test designers need to approach a set of rater training procedures to ensure raters' scoring reliability. This process is to be cyclical with inter-rater and intra-rater reliability steps to detect potential inconsistencies (Bachman & Palmer, 2010; Luoma, 2004; Wang, 2010; Weigle, 1998). In an earlier study, Wigglesworth (1993) explored the effect of rater feedback on eight raters' rating performances as regards to rater bias issue. Rasch analysis was utilized to examine inter-rater and intra-rater reliability. Firstly, these eight raters joined in a rater training program with two steps: a) an individual session, and b) a group session. In the first session, the raters formed 'assessment maps' including items about bias analysis. Furthermore, they received detailed feedback on their rating patterns on these maps. As for the group session, they firstly evaluated some tasks in groups, and then they worked on the test tasks by themselves. Following all the feedback and training sessions, the ratings were analyzed through FACETS. The findings showed that the feedback sessions were fruitful in terms of the severity and leniency of assigned scores and at the same time these sessions decreased the effect of rater bias. Similar to the methodology of the previous study, Lumley and McNamara (1995) utilized Rasch analysis to explore the raters' consistency and rater bias along two separate rater training programs. The data of this study were gathered from the speaking section of OET, which is a kind of test developed for medical employees working in Australia. The research disclosed that feedback reports could be used as an effective technique while addressing the severity and leniency of the assigned scores.

Papajohn (2002) investigated raters' reasoning patterns following a rater training program. The exam data were gathered from SPEAK, which is an unused version of Test of Spoken English (TSE). Four expert raters as evaluators and 9 trainee raters contributed to the study. Firstly, the raters completed the rating training steps (2 hours individual and 10 hours group work). Secondly, they were required to form a concept map following each rating. In addition, they were asked to make remarks about the concept maps they formed. Thirdly, relying on a set of rating principles, the evaluators examined and assessed each

concept map by looking at the similarities and differences. The examination of concept maps revealed that the raters continued their personal beliefs and attitudes towards the rating process although they all had the same rating training content. Furthermore, rater assessors observed some variation of rating patterns across the trainees. As for a methodological implication, concept mapping technique could be used as an alternative to verbal protocols.

Using an experimental research design, Stitt et al. (2003) conducted a study focusing on 'evaluation fidelity' within two different studies: setting 1 (between teacher and teacher) and setting 2 (between teacher and students). In study 1, the researchers aimed to explore whether there would be a significant difference in raters' inter-rater reliability after the rater training. A total number of 19 raters joined in the rater training program, which included reviewing assessment criteria, examining the feedback reports and sample performances. According to independent samples *t*-test findings, thanks to this program, raters assigned the speaking performances more consistently. In study 2, the experimental group raters who instructed their students how to assess spoken samples via criterion-based principles scored more homogeneously than the control group raters who did not.

Xi and Mollaun (2011) carried out an extensive study exploring how well raters from India could award scores to test takers with different L1 backgrounds as well as examining the adequacy of the rater training program. The speaking part of the TOEFL iBT test was utilized as exam data. Given the importance of high stakes tests in test takers' lives, rater recruitment and raters' professional development processes are managed seriously and strictly to ensure reliability and validity of raters. As such, the researchers followed a similar procedure as TOEFL iBT does. Twenty-six raters from India were equally split as rater group 1 (regular training group) and rater group 2 (special training group). In session 1, both groups received the regular training program and filled out a rater feedback survey. However, in session 2, only rater group 2 received the special training and filled out the second feedback survey. These feedback surveys aimed to reveal raters' thoughts about the efficiency of the training program and determine the difficulties that they faced while scoring Indian test takers' performances. The overall quantitative findings of this study showed that a detailed and well-planned rater selection and training program decreased the effect of rater bias towards certain L1 accents. Thanks to the effectiveness and practicality of rater training

sessions, raters from India were as successful as expert TOEFL iBT raters. The second rater group covered topics related to rater bias and various accents in the special training program, resulting in more homogeneous scores across the groups. Finally, according to the findings retrieved from the surveys, the raters reported that the training programs were effective and practical in terms of rating consistency and building affective skills. However, they suggested that more comprehensive instruction of Indian test takers' pronunciation and intonation samples could have been covered in the training programs.

In a study examining raters' consistency and rater performance, Yan (2014) found that regular rater training sessions brought considerable benefits to the raters. For instance, the raters received training assigned more consistent scores and approached the scale more systematically than the untrained ones. The speaking performance data were collected from the Oral Proficiency Test, which evaluated the foreign teaching assistants' communication abilities of English. 253 test takers and 11 raters of OEPT contributed to the study. Similarly, Kang et al. (2019) revealed that an effective rating training program improved the rating quality of novice and experienced raters. 28 test takers' speaking performances, rater training program essentials, and benchmark scores were taken from Educational Testing Service (ETS). A total number of 40 raters participated in the website based rater program prepared by ETS. They regularly completed the training steps and compared their scores with the benchmark score given by expert ETS raters. Following this rater training, as regards to severity and leniency of the scores, the raters who had given higher and lower scores became more similar to each other. In fact, thanks to the sample speaking performances showing the test rubric's details and benchmark scores, the raters became less subjective and were more confident while using the test scale.

To summarize, the use of feedback channels such as concept maps, assessment maps, and group interaction within the rater training framework impact L2 speaking assessment positively (Lumley & McNamara, 1995; Papajohn, 2002; Wigglesworth, 1993). The experimental research design can give new insights into the effectiveness of rater training programs (Stitt et al., 2003). In addition, rater training programs need to operate compactly since each step from rater selection to session frequency requires a comprehensive approach (Kang et al., 2019; Xi & Mollaun, 2011; Yan, 2014).

2.3.4. Rating Scales

Creating a speaking test scale requires a detailed study of numerous aspects such as test purposes, test audience, construct definitions, scale descriptors, and scale type. In fact, rating scale development is not merely compiling descriptors or anchors from various reliable and valid scales. On the contrary, it is a process through which test designers must adopt the principle of suitable rating building method(s): 'intuitive', 'qualitative', and 'quantitative'. Varying methods will definitely bring a wider perspective, contributing to the practicality and effectiveness of rubrics (Kim, 2006; Luoma, 2004). L2 speaking assessment is relatively more challenging than the other assessment types due to the nature of speaking skills. Above all, it is more difficult for raters to pay attention to test takers' speech patterns in a limited time period because raters might be affected by a lot of factors such as accents, gestures, and emotional states (Winke, 2012). Therefore, referring to the underlying constructs, test designers are to create better harmony in the development of rating scales. While doing so, they need to start rating scale and speaking test task development at the same time as these two are interrelated (Fulcher & Davidson, 2007). As for the framework of rating scales, three main categories can be provided: a) 'orientation (user, assessor, and constructor)', b) 'scoring (analytic and holistic approaches)', and c) 'focus (real world and construct)'. The assumption that rating scales have the purposes of making the assessment process more objective and practical naturally separates them into certain types. For instance, test taker friendly scales illustrate information about the target levels while rater friendly scales provide clear and concise points regarding the test constructs so that raters can make use of them during the test interviews. Test management friendly scales show comprehensive and detailed samples of speaking test tasks (Fulcher, 2014; Luoma, 2004; North, 2012).

No doubt there is not a perfect rating scale type, namely, each rating scale type has its own advantages and disadvantages (Chalhoub-Deville, 1995). The comparison of different types of scales can give us the idea of a starting point in the process of rating scale development. To illustrate, test designers would most likely choose a generic scale rather than a task scale due to operational and technical reasons. While task scales target specific

skills about the task, generic scales are good at assessing the overall constructs by focusing on general skills. Similar rating scale comparisons can be made as holistic and analytic scales or behavioral and real world scales (Green, 2014; Luoma, 2004). Holistic scales are known to provide a quick and practical assessment opportunity for raters and at the same time their descriptors are rater-friendly in terms of interpretation and memorization. As for the drawbacks, the descriptors might not reveal the variations of scoring across levels. Also, giving feedback to test takers may not be possible due to its overall scoring nature (Brown, 2004; Fulcher, 2014; Weir, 2005). Unlike holistic scales, analytic scales can provide a more comprehensive scaling range to raters and also ample information of test takers' strengths and weaknesses. The fact that analytic scales have numerous subcategories with detailed descriptors might not be very practical for raters especially in performance assessment. Another disadvantage of analytic scales is that they may limit raters, resulting in bias (Alderson et al., 1995; Brown, 2004; Luoma, 2004; Madsen, 1983).

Brown (2006) investigated how six IELTS raters' reviewed and used the analytic scale designed for the speaking exam of IELTS. The spoken data were gathered from 12 test takers' performances whose scores ranged from 5 to 8 out of 9 points. Having received the verbal report training, the raters provided these reports through stimulated recall methodology. The researcher coded the reports according to the main categories of the scale. After carrying out the inter-coder reliability, the researcher analyzed the coded items based on the research questions. In addition, the raters responded to a questionnaire to corroborate the findings from the verbal reports. The research questions focused on three main aspects: a) how the raters interpreted scales, b) how they differentiated the levels in the scale, and c) the difficulties that they faced while grading. The findings of the study showed that the raters thought that this analytic scale was straightforward and practical although they addressed some issues such as some complexities in the fluency and the ambiguities in the lexical parts. In addition, the raters reported that it was difficult to distinguish test takers' level in the fluency and coherence category of the scale. They also had problems with distinction of some words in the scale. Despite some of the drawbacks, the raters generally seemed to be in conformity with this analytic scale. Furthermore, what makes the finding of this research valuable is that Brown (2007) had explored the effectiveness of a holistic scale having been used in IELTS speaking tests. The researcher reached striking findings that the raters showed

variations in terms of complying with the holistic scale. While some of the raters covered each descriptor in the scale, others interpreted the scale less than expected. In fact, the holistic scale was found ineffective due to the ambiguities in the descriptors.

Using a mixed method research design, Chuang (2009) conducted an extensive study to explore the raters' score variation across holistic and analytic scales as well as their opinions about the categories of both scales. While the quantitative data were collected from the rater questionnaire and the holistic and analytic scales, the qualitative data were gathered from the rater interviews. 62 raters were selected from the university instructors working in one part of Taiwan. As for the findings, inferential statistics results showed that there were not any significant differences between the holistic and analytic scores. Secondly, the raters ranked the subcategories of analytic scale in the order of importance and most of them determined 'comprehensibility' and 'pronunciation' as the two most important ones, respectively. Given the least important components, 'vocabulary/word choice' was at the top of the list. These findings might show the areas on which the raters labelled their scoring reasons. The third research question investigated whether there was a relationship between the raters' age and the score variation on the use of holistic scale. The researcher opted to use the holistic scale because there were not any significant differences between the holistic and analytic scales. According to the findings of the third research question, the raters who were in the age group of 21 and 30 awarded lower scores than the other age groups. However, there were not any significant differences between the raters' training background and the score variation and similarly between raters' experience and the assigned scores. Finally, the researcher underlined the importance of educational needs while determining whether to use either holistic or analytic scale in speaking assessment.

Isaacs and Thomson (2013) explored whether there were any relationships between rating scale length and rater experience as regards to score variation and 5-point or 9-point scale preference. Speaking performance patterns were collected from 38 ESL students with different L1 backgrounds. A total number of 40 experienced and inexperienced raters contributed to the study as subjects. The experienced ones were the graduates of language-teaching related departments. However, the novice ones were graduate students of non-language teaching related departments. Utilizing mixed method research design, the

researchers created an experimental design by assigning 5-point and 9-point scales to the raters from each experience group, verbal protocols, and post-interviews. According to the quantitative findings of rating scale length, there were not any significant differences in raters' score variation. However, the qualitative findings revealed that the raters reported some issues about the scoring length accent in the scale. While some of the raters expressed their discomfort with the use of the 5-point scale, most of them were not glad about the scoring length of the 9-point scale.

Reviewing the problems with traditional scales, Upshur and Turner (1995) explored the effectiveness of a novel scale assessing grammatical and communicative components. This scale was called 'empirically derived, binary-choice, boundary-definition (EBB)'. As for the rating scale development process, the research team worked on eight test takers' speech samples by separating them into two groups (4 weak and 4 strong performances). Finally, a set of double questions with numerical ratings were determined to differentiate the levels. 99 pupils with L1 French background from Canada completed the story-retelling test, which were recorded for the use of rating. After that, two raters assigned these performances via EBB scale. The findings of the study revealed that a high reliability and validity can be ensured since EBB scales are based on the analysis of actual performances of test takers. Another strength of this scale is that they are rater-friendly thanks to a set of double questions and an easy scaling system. Overall, the researchers anticipated that performance-based scales like EBB can be an alternative to standard test scales in speaking assessment.

Similar to the concerns about standard test scales expressed by Upshur and Turner (1995), evaluating the details of the measurement-driven approach and performance data-driven approach, Fulcher et al. (2010) designed a scoring instrument called Performance Decision Tree. The researchers were critical of measurement-driven approaches due to the contextual issues. Giving CEFR as a typical example of this approach, the researchers expressed that the levels in the CEFR scale lack performance patterns retrieved from actual learners. As such, they suggested that scale designers could form a performance data-driven scale by analyzing the discourse and pragmatic aspects of the collected real life interactions. With this in mind, this instrument was created as a model of performance scale with three main components and 10 sub-components: a) 'discourse competence', b) 'competence in

discourse management', and c) 'pragmatic competence'. Although this type of scale might be challenging for raters in terms of timing and practicality, it can provide numerous chances of classroom feedback following speaking tests. Referencing the L2 acquisition findings, the researchers stressed the need for more contextual and interactional based scales to describe authentic learner performance patterns.

To sum up, regardless of the type of rating methods, it can be said that rating scales naturally have their own advantages and disadvantages. Therefore, a comprehensive situation analysis is necessary before designing or implementing a rating scale (Chalhoub-Deville, 1995; Green, 2014; Madsen, 1983). While some of the studies revealed the merits and demerits of analytic and holistic scales in L2 speaking assessment (Brown, 2007; Brown, 2006; Chuang, 2009), others investigated rating scale length (Isaacs & Thomson, 2013) and the performance-based scaling systems (Fulcher et al., 2010; Upshur & Turner, 1995).

This dissertation did not aim to explore the effects of different rating scales on raters' scoring patterns or variation. Given the certain advantages of analytic scales for speaking raters, the institution where the main data were collected opted to design an institutional analytic rating scale. Furthermore, the scale designers in that institution added a strengths and weaknesses section under the main descriptors, aiming to provide feedback and data to learners and raters. Namely, this scale was not a directly adopted rating scale from other institutions or publishing companies. Since the raters were already familiar with using this scale, the researcher decided to utilize the same scale for this study without making any changes.

2.4. Rater's Rating Experience in L2 Performance Assessment

Given the dictionary definition, an experienced person is called someone "possessing skills or knowledge because s/he has done something often or for a long time" (Longman, n.d., para. 1) According to this basic definition, an experienced rater can be classified as the one who has had a great number of and subsequent rating duties for a period of time. However, the question is whether we should call a rater with only a lot of rating practices

experienced or not. Naturally, this basic dictionary explanation will not be adequate to explain who an inexperienced or experienced rater is. For instance, Kim (2015) defines the rater experience groups in the study based on a set of detailed experience criteria. Namely, the length of time spent on rating performance assessment might not be the major determining factor in terms of defining rater experience. In other words, it might not guarantee the level of experience. Within the framework of performance assessment, rater experience seems to be related to numerous factors such as raters' background, research context, exam type, rater training, and teaching experience (Bonk & Ockey, 2003; Cumming, 1990; Davis, 2016; Fox, 2003; Weigle, 1998). Comparing the number of studies investigating the effects of rater experience on score variation and rating behaviors, there are many studies conducted in L2 writing assessment research (Barkaoui, 2010a; Cumming 1990; Delaruelle, 1997; Leckie & Baird, 2011; Lim, 2011; Myford et al., 1996; Sakyi, 2003; Song & Caruso, 1996; Wolfe et al., 1998), only few studies have examined the effects of rater experience in the context of L2 speaking assessment (Bonk & Ockey, 2003; Davis, 2016; Kim, 2015).

The studies investigating the effects of rater experience in the context of writing assessment have revealed varying findings. To illustrate, Cumming (1990) found that six expert and seven novice raters showed different rating patterns while evaluating 12 essays in terms of utilizing the components of the scale. While the novice raters gave higher scores, the expert raters tended to award more consistent scores. In addition to the quantitative findings, the researcher observed a set of decision-making strategies employed by novice and expert raters. Collecting the data via think-aloud protocols, Delaruelle (1997) investigated the impact of raters' teaching experience on rating patterns and score variation. Three experienced and inexperienced teachers awarded scores in a test that they did not have any experience in beforehand. The overall results showed that experienced teachers were better at commenting strategies while evaluating than the inexperienced teachers.

Both Cumming (1990) and Delaruelle (1997) revealed that experienced raters generally performed more consistently than the inexperienced ones. Similarly, Sakyi (2003) examined five inexperienced and experienced teachers regarding their rating behaviors and scoring variation. Displaying some differences between two rater experience groups, the

researcher revealed that experienced raters employed more effective and quick strategies while using the scale than the inexperienced ones. Lim (2011) explored whether there would be any changes in experienced and novice raters' rating tendencies in three-month subsequent rating process. According to the findings, the experienced raters did not show any scoring differences throughout the process, namely, they were more consistent while scoring. Given the effectiveness of the added rating practice, novice raters made considerable progress as regards to their rating quality.

Barkaoui (2010a) researched 31 inexperienced and 29 experienced raters' scoring variation regarding analytic and holistic scale scoring methods. The study revealed that inexperienced raters assigned lower scores when utilizing analytic and holistic scales, and at the same time the inexperienced group were more inconsistent in terms of severity and leniency of the scores. As for the experienced raters, they tended to award lower scores to the test takers while using the analytic scale. In overall, the findings addressed inconsistencies as to rater experience and reliability. However, some of the studies revealed no direct relationship between the rater experience and score variation. For instance, Song and Caruso (1996) found that there were not any considerable differences between experienced and medium-experienced raters' analytic scoring while there were some variations in their holistically assigned scores. Leckie and Baird (2011) conducted a study in which three rater groups (supervisor raters, novice raters, and experienced raters) scored a nationwide English writing test for secondary school students. The results revealed that there was not any significant difference between experienced and novice raters in terms of severity and leniency of the assigned scores. However, the supervisor rater group showed a different rating pattern compared to the experienced and novice ones.

Unlike the studies in writing assessment research, relatively fewer studies have examined the impact of rater experience on rating patterns and scoring variation. Davis (2016) investigated the impact of experience and training on 20 experienced English instructors' rating consistency and patterns in TOEFL iBT speaking test context. Despite these instructors' high-level of experience in teaching English, they were labelled as novice speaking raters due to their limited experience of rating speech samples. Given the number of tasks 1 and 2 in the speaking test, the raters assigned scores to 480 speaking performances

gathered from 240 test takers. While grading, the raters used the TOEFL iBT holistic speaking rubric, which was modified by the researcher in the format of six-point scale instead of four-point scale. Within the framework of the study, the raters firstly completed the orientation program and then they assessed 100 speaking performances in the scoring session 1. This session was followed by a rater training session in which the raters had the benefit of the rubrics, scoring exemplars, and calibration sessions. After that, the raters scored 100 speaking performances in three subsequent scoring sessions. The ultimate aim of this study was to observe whether there would be changes in the low-experienced raters' scoring variation thanks to the experience gained by the training program. The results of the study showed that there was a moderate impact of these gradual training and scoring sessions on the raters' scoring variation. This finding might mean that the raters had been somewhat less reliable and consistent before the training sessions. Given the limitations of this study, the underlying reasons why the raters in this study had a certain level of rater reliability before the training were ambiguous. However, the researcher assumed that the raters' level of experience in teaching English might have a positive effect on their rating consistency and accuracy. This actually could give new insights into what type of components should be added while defining a low-experienced and high-inexperienced rater.

Bonk and Ockey (2003) conducted a study to explore the effects of various facets such as prompt, raters, rating scales, and test items on score variation. This study was based on a group speaking test conducted at a major university in Japan. In this test, test takers firstly completed a couple of integrated activities such as an orientation video and video-based reading, vocabulary and grammar tests. Following that, they carried out a group speaking task in the format of discussion. Utilizing a scale with five components, two independent raters assigned scores to the test takers who were varying degrees of English language related department students. The data were gathered from two successive test sessions called 'admin1' and 'admin2' to explore the rater variation across different time periods. A total number of 20 and 26 raters contributed to these two sessions, respectively. Being graduates of English language related or teaching departments, the raters received an intensive rating training and norming sessions. The findings of this study revealed that newly recruited raters assigned more inconsistent scores than the experienced ones. However, these new raters did not show inconsistencies in the next year test session. There were to some

extent variations among the raters across two test sessions despite the added experience of training aid. Therefore, the researchers underlined the importance of decreasing the effect of score variation stemming from rater differences, which could cause more issues than other facets.

Utilizing a qualitative research design, Kim (2015) explored the rating patterns of three varying rater groups: a) novice raters, b) developing raters, and c) experienced raters. The research was carried out in a language school situated at a major university in the US. A total number of 18 test takers provided 6 speaking performances from each proficiency level: advanced, intermediate, and beginner. Three raters from each rater group contributed to the study. As for the categorization of rater groups, novice raters were selected from the language related department graduates who had had no teaching English or rating performance assessment experience. Developing raters were the graduates who had some teaching English and rating experience. Experienced raters were the graduates with more than five year teaching English and rating experience. In addition to this, experienced raters had a profound rating experience in the speaking placement test, from which the speaking performance data were collected. Completing a rater's background questionnaire, the raters were provided with an analytic scoring rubric with 5 sub-components. Utilizing this scale, the raters participated in three rating sessions with one-month apart. Given the qualitative nature of this study, the main data were collected through verbal reports to compare three varying rater experience groups' rating patterns. The overall findings of this study revealed that there were considerable differences across three rater groups especially in novice and developing raters' rating tendencies. Looking at the verbal report findings in detail, the researcher observed that the novice and developing raters had similar difficulties in assigning the suitable level, internalizing the rubric components, and focusing on the scale descriptors in a balanced way. However, after completing the third rating session, these two groups made substantial progress in the areas that they had problems with. Unlike both novice and developing raters, the experienced raters were not observed to show such inconsistencies and confusion in terms of using the scale components and assigning appropriate proficiency levels.

All in all, some of the studies found that experienced raters showed more consistent rating tendencies than inexperienced ones (Cumming, 1990; Delaruelle, 1997, Lim, 2011). Furthermore, others explored whether there would be any differences in rater experience groups' rating patterns and scoring variation regarding holistic or analytic scale use (Barkaoui, 2010a; Song & Caruso, 1996). However, some of the studies did not observe any differences between experienced and inexperienced raters (Leckie & Baird, 2011; Song & Caruso, 1996). All these studies were within the framework of writing assessment research. As for the rater experience effects in speaking assessment, the studies focused on the relationships between added rating practice support and rater experience background, implying that less experienced raters improved their rating performance in the course of time (Bonk & Ockey, 2003; Davis, 2016; Kim, 2015).

In this present dissertation, considering the niche in the literature, three rater groups (low-experienced, medium-experienced, and high-experienced) were formed by utilizing a rater experience scale to define the boundaries of the rater groups. Using the quantitative data set, this study initially compared the analytic scores of low-, medium- and high- quality L2 speaking performances. Secondly, the study explored whether there are any significant differences among the analytic scores assigned by low-, medium- and high experienced raters.

2.5. Speaking Performance and Text Quality in L2 Performance Assessment

Performance assessment consists of numerous factors that have the potential of affecting one another. Therefore, the interaction across these factors is worthwhile researching in terms of the fairness of the testing process (Bachman & Palmer, 2010; Brown, 2004; McNamara, 1996, Weigle, 2002). Although varying quality speaking performances in L2 speaking assessment is an issue waiting to be explored in detail, speaking performance is interestingly one of the under-researched aspects. However, unlike speaking assessment, there are a lot of studies investigating the impact of essay quality on writing scores (Brown, 1991; Daly & Dickson-Markman, 1982; Engber, 1995; Freedman, 1981; Han, 2017; Şahan, 2019).

Brown (1991) conducted a study to compare the differences between ESL and NES university students' writing scores assessed by eight raters from ESL and English departments at the University of Hawaii. A total number of 56 essays, which were randomly selected from the whole set, were gathered from each of the student groups. An institutional holistic rubric with 6-point scale was utilized while rating the essays. The statistical results showed that there were not significant differences in the ESL and NES students' essay scores. Similarly, there were not any considerable differences across the raters from ESL and English departments. Additionally, the researcher required the raters to rate the strong and weak aspects of the essays in terms of cohesion, mechanics, content, lexicality and syntax. For instance, both rater groups determined content as a positive aspect and syntax as a negative one. All in all, the study revealed that the rater groups approached the varying quality essays differently regarding the aforementioned points.

Having mentioned before, Brown (1991) found no score variation across the different groups of raters and quality of essays despite some rating variations. As for the text quality, some of the studies aimed to examine the relative effects of previous scoring expectation on the rating process. To illustrate, Daly and Dickson-Markman (1982) explored the contrast impact of varying quality essays on the rating process. Designing two steps in this study, the researchers primarily aimed to categorize the essays in three distinct qualities: low, moderate, and high. A total number of 37 teachers with writing rating experience classified these essays. Secondly, utilizing the experimental research design, the researchers asked 157 raters to assess the compiled essays in the format of varying quality series in order to examine the impact of previous essay quality. In total, six rating scenarios were formed and one essay was set as neutral. For instance, while the raters evaluated four high-quality essays following neutral in the first round, they assessed four low-quality essays and one neutral at the end. The results of this study indicated that there were significant differences between the aforementioned round 1 and round 2, meaning that moderate level essays were evaluated as low-quality due to the effect of high-quality essays. However, the same quality of essays was given higher scores following the low-quality essays.

Freedman (1981) carried out a study to investigate the impact of various factors such as essays, raters, and environment on the assigned holistic scores. The researcher required test takers to opt for one essay topic across eight topics and write in 45 minutes. Four raters, selected from the doctoral candidates of the English literature department at Stanford University, participated in the study. Before each rating session, these raters were trained to ensure the rater reliability. The results of the Cronbach's alpha signified a high level agreement across the raters. During the rater training, two trainers led different formats of sessions. The first trainer focused on the topical discussions. On the other hand, the second trainer organized rating discussions through which the raters mentioned their beliefs regarding the characteristics of good and bad essays. Furthermore, the trainer informed the raters that the test takers were limited to a maximum of 45 minutes. As such, the researcher assumed that the raters tended to be more lenient especially following the second trainer's session. In fact, the score of a low and high-quality essay showed variations based on the level of expectation.

Engber (1995) examined whether raters were affected by lexical aspects in varying quality essays. Sixty-six essays were collected from a placement exam organized at the ELT program of Indiana University. The proficiency level of test takers was in level 4 out of 7 levels ranging from intermediate to high-intermediate. Utilizing a holistic rubric with 6 point-scale adapted from a high-stakes test, 10 raters, who had experience in high-stakes tests, awarded scores to the test takers' performances. As for the reliability of essay scoring, the researcher conducted the Pearson coefficient test, resulting in a good level of inter-rater reliability (r = .93). Investigating the lexical variation (with and without error), lexical error percentage, and lexical error density, this study revealed that there was a significant difference regarding lexical error percentage. This finding actually meant that the raters tended to give lower scores to the essay with more lexical errors.

Utilizing a mixed-method research design, Han (2017) investigated the effect of low-, medium-, and high-quality essays on the score variation and rating behaviors. The essays were collected from the EFL learners studying at three different universities in Türkiye. Following the essay quality division process, 30 essays were determined as the master data. A total number of five raters with similar backgrounds and experience level contributed to

the study. A holistic rubric with 10-point scale was used. As for the qualitative aspect of the study, the raters recorded their thoughts while assessing six essays following the procedure of think-aloud protocols. The findings of this study revealed that there was more variation in the assigned scores of the low and medium-quality essays than high-quality ones. Similarly, G-theory analyses showed that the raters scored the high-quality essays more consistently. In addition, G-theory analyses disclosed that essay quality and raters were two top sources of variation, 37.2 and 24.8 percent respectively.

Similarly, Şahan (2019) conducted an extensive study in which the researcher aimed to explore score variation related to low- and high-quality EFL essays assigned by three distinct rater groups. A total number of 33 raters contributed to the study. While 15 of the participants were EFL instructors working at the same school of foreign languages in a state university, 18 raters were language teaching professionals from various universities across Türkiye. These participants were classified as low-experienced, medium-experienced, and high-experienced raters. A total number of 104 essays were collected from the department of ELT in a state university in Türkiye. Following the essay quality division process by expert raters, the researcher utilized 25 low- and 25 high-quality essays for the main study. This research relied on descriptive and inferential statistics as well as G-study. The results of the inferential statistics analysis showed that there were significant differences in the scores assigned to low-quality and high-quality essays. As for understanding the correlation between the rater experience and the varying essay quality, a couple of non-parametric tests were conducted. The findings showed that while there were not any significant differences in the scores awarded to high-quality essays, there were significant differences assigned to low-quality essays. Furthermore, the subcategories of the rubric were examined. The results showed that there were significant differences in the scores assigned to the mechanics category of low-quality essays. However, there were not any significant differences in the other subcategories. G-studies were conducted for mixed- and separate (low-and high-) quality essays. Given the mixed-quality findings, the largest variance component was due to students (45.3%) and the second biggest factor was residual component (31.3%), which corresponds to essays, raters, essay quality and other factors. The G-study results for low- and high-quality essays showed that residual was the biggest variance component for low-quality (46.2%) and high-quality (54.2%) essays. In addition, the researcher observed more

consistent scores in the mixed-quality and low-quality essays. Considering all these points, this research provides valuable implications for establishing a novel rater training model, pairing the raters in a double-grading system, revisiting scoring rubrics, and designing essay tasks.

In light of the findings of the previous studies regarding the relationship across essay quality, assigned scores, and rating patterns, it can be claimed that different quality essays might have an impact on the aspects of performance assessment. We can observe more traditional research methods and approaches in earlier studies having examined essay quality (Brown, 1991; Daly & Dickson-Markman, 1982; Engber, 1995; Freedman, 1981). However, both Han (2017) and Şahan (2019) utilized descriptive and inferential statistics as well as G-theory framework so as to explore whether essay quality had an impact on score reliability and relationship with other sources of error. Given that speaking performances quality has not been researched extensively, this study examined L2 speaking performances of three qualities to reveal whether low-, medium-, and high-experienced raters would show different rating patterns while assessing low-, medium-, and high-quality speaking performances.

2.6. Rater Cognition and Decision Making in L2 Speaking Assessment

Rater cognition is related to "the mental processes occurring during scoring, at either a conscious or unconscious level" (Davis, 2012, p. 9). Considering this definition within the scope of rater cognition research, the characteristics that raters possess while scoring and raters' decision making process are two focal centers of interest (Bejar, 2012). Namely, raters may give their scores based on either deliberate or unintentional tendencies. Even if raters with varying experience can award very similar scores, the ways in their decision making patterns may show differences. Therefore, rater cognition based studies in language assessment are highly crucial to understand raters' decision mindset. Given the weight of the rater cognition studies in L2 performance assessment, it would not be wrong to claim that L2 writing assessment studies have relatively outnumbered the ones in L2 speaking assessment (Baker 2012; Barkaoui, 2007, 2010b; Cumming et al., 2002; Gebril & Plakans, 2014; Han, 2017; Şahan & Razı, 2020; Vaughan, 1991; Wolfe, 2005). This might be due to the availability of various rater cognition models and approaches in the framework of L2

writing assessment. As to major rater cognition models in writing assessment, I can mention four crucial studies (Frederiksen, 1992; Freedman & Calfee, 1983; Lumley, 2005; Wolfe & Feltovich, 1994). Despite having some basic differences, these models complemented each other as regards to theoretical and methodological aspects. Naturally, they have contributed to the development of numerous studies in L2 writing assessment field. However, there are not any similar models for the L2 speaking assessment context. In this regard, the current situation has raised concerns whether those writing models can be used in speaking assessment studies. In essence, without considering the nature and dynamics of L2 speaking assessment, adopting writing models directly might not serve the purposes of understanding decision making strategies that speaking test raters use. Nonetheless, those models may assist speaking assessment researchers to conceptualize the basis of rater decision making patterns (Davis, 2012).

There are two fundamental questions giving direction to rater cognition studies: a) In what ways do raters show variations in their rating behaviors? And b) What are the reasons for these rating differences? When the majority of rater cognition studies (either writing or speaking assessment ones) are examined in detail, these two main research questions have been the dominant motives for scholars. However, it would be more comprehensive if scholars utilized cognitive processing theories while researching raters' thinking processes (Han, 2016). As for rater cognition studies in speaking assessment context, some of the studies examined the differences that raters showed whilst evaluating the same responses (Ang-Aw & Goh, 2011; Orr, 2002) while others focused on the rater cognition and decision making patterns regarding rater background and characteristics (Cai, 2015; Chalhoub-Deville, 1995; Gui, 2012). One of the studies specifically investigated the areas that raters mostly highlighted and agreed on (Brown et al., 2005). Another study explored the rating approaches that raters adopted while assigning scores to test takers' performances (Pollitt & Murray, 1996).

Using both quantitative and qualitative research methods, Ang-Aw and Goh (2011) examined raters' decision-making process in the speaking section of a high-stakes test conducted in Singapore. This test is called the 'O' level English examination including writing, reading, and speaking parts. Given the importance of this exam, high-school

students have to take an acceptable score to have access to higher education institutions in Singapore. As for the participants, seven raters that were experienced in rating the exam and four secondary school students contributed to the study. The overall results of this study revealed that the raters showed certain differences regarding the use of rubric descriptors, personal rating beliefs and principles, and rating approaches. 'Personal response', 'clarity of expressions', 'engagement in conversation' were the major headings in the rubric. While some of the raters only focused on the candidates' conversation skills, others opted to favor the quality of speaking performances. This variation naturally had the potential of affecting the reliability of the scores in that high-stakes examination. In addition to the factors in the rubric, the raters relied on their own personal beliefs such as the originality of ideas, lexical richness, the relative impact of other test takers' performance, and test takers' attention. In fact, the researcher assumed that raters' varying perception of speaking proficiency and assigned scores could be related to the personalized interpretation of exam constructs. Finally, examining rating approaches, this research found that while the majority of the raters adopted 'synthetic approach', some of them relied on 'mixed approach'. Only one of them tended to use an 'objective approach'. In addition, the study revealed some differences across the raters regarding two issues: severity/leniency and comparing test-takers' performances with each other.

In a similar vein, Orr (2002) explored the decision-making process that 32 experienced raters underwent while assessing the speaking part of the test called First Certificate in English. To collect the data, the raters were informed about the principles and objectives of verbal reports. The qualitative data were utilized to corroborate the data yielded from the assigned scores. As for showing the nature of the score and its reasoning variation, the researcher observed that even if two different raters assigned the same score, their reasoning behind the scores were totally varied. While one of the raters focused on the frequency of errors, the other considered the grammatical aspects. The raters mostly focused on three distinguishing features in test takers' performances: a) test takers' self-presentation skills, b) raters' inter-comparison of test takers' performances, and c) assigning scores based on test takers' general performance indicators. One of the most crucial aspects of this study was that the raters tended to assign scores apart from the rubric criteria. The raters reported that the ambiguous and incomplete parts in the speaking test rubrics might have forced them

to adopt independent ways while assessing speaking performances. Given the importance of rater cognition studies in L2 speaking assessment, this study actually gave profound implications for the development of speaking test rubrics and rater training.

Brown et al. (2005) conducted two studies in which they explored both rater cognition and speech samples, respectively. In Study 1, the main data were collected from the verbal reports created by the total number of ten experienced raters. As for the data analysis, the researchers firstly transcribed the verbal reports to conduct data segmentation and draft coding. Following all these processes, the researcher and one independent researcher worked collaboratively to compute inter-coder kappa analysis, resulting in a high level of interrater reliability. The findings of the first study revealed that the raters focused on four areas: 'content', 'phonology', 'fluency', and 'linguistic resources'. When each of these main headings were analyzed in detail, the researchers determined particular subcategories. The raters evaluated aspects such as hesitation, fillers, pauses, speech rate, and repetition under the heading of fluency category while they tended to rate test takers' task fulfillment, framing, and ideas in the content category. As for the linguistic category, they highlighted four subcategories: grammar, academic/daily expressions, and textualization. Pronunciation, intonation, stress, and rhythm were the focused areas in the category of phonology. In addition, this study examined the points specified by the raters under the subcategories. For instance, the raters featured introduction and conclusion aspects within the category of framing. Unlike the results of the previous rater cognition studies, the content category was the most salient finding of this study. Given the importance of completing the task content in high stakes speaking tests, it would not be wrong to claim that content needs to be taken into consideration in L2 speaking assessment. Another important finding of this study was that the raters did not create a separate category regarding discourse. Instead of doing this, they mentioned discourse features under other main categories. The second research question in this study explored the percentages of rater comments across five proficiency levels. The linguistic resources and content were two top frequently commented categories. The percentage of comments tended to rise in the higher levels of proficiency. However, the percentages showed a downward trend only in phonology and fluency categories as the level of proficiency increased. Even if the raters made comments in all categories across five different levels, they made fewer comments regarding discourse

and framing as the levels decreased. Furthermore, the researchers observed that the raters approached the main categories differently in terms of independent and integrated tasks.

Utilizing regression analyses and verbal report methodology, Cai (2015) examined rater cognition as to varying rater types' rating tendencies in a high stakes speaking test. Thanks to a holistic judgment task, the researcher categorized the participating 126 raters as three distinct rater types: balanced, form-oriented and content oriented raters. During this process, the raters were required to assign scores based on five criteria: 'pronunciation', 'grammar-vocabulary', 'organization', 'richness in content', and 'topic relevance'. Test scores were retrieved from a high stakes test called 'Test for English Majors (TEM4-Oral)' conducted in China. Although this test had tasks such as topic-based talk, story-retelling, and debate-based small talk, the researcher opted to use topic-based talk due to its commonality. The raters received a short training program designed for the test. According to the findings of the first research question, there were significant differences in the mean scores of five criteria in three rater groups. For instance, balanced raters focused on topic relevance and content. However, the content-oriented raters gave importance to topic relevance and the form-orientated raters tended to favor pronunciation and grammar-vocabulary while deciding on their scores. As for the second research question, the researcher explored the level of weight that the raters gave each criteria in the rubric and the percentages of rater opinions. Pronunciation was the most important aspect for each rater in the form-oriented group. Relevance and richness of content were two most favorable criteria for the majority of balanced raters group. Relevance and richness of content were prioritized by the content-oriented raters. In terms of the percentages, most of the comments were made on content by the balanced and content-oriented raters. However, a high percentage of comments on pronunciation was observed for the form-oriented raters. In the final research question, the researcher examined whether there were any differences in three rater groups regarding validity. The verbal reports revealed that the form-oriented raters attached more significance to form criteria and less importance to richness in content. In addition, the aspects of form were dominant while these raters were evaluating the content. However, balanced and content-based raters tended to assess these two criteria separately.

Chalhoub-Deville (1995) conducted a study to explore the aspects to which varying rater groups gave importance while assessing speaking test samples provided by six participants. A total number of 82 raters, whose L1 was Arabic, from three main groups contributed to the study. The first group included the raters who were native Arabic speaking teachers of Arabic as a Foreign Language (AFL) in the US and the second ones were formed from the native Arabic speaking people living in the US more than a year. The last group corresponded to the native Arabic speaking people living in Lebanon. The second and third group raters were not instructors, but university students in their resident countries. Firstly, thanks to multidimensional scaling regressions analyses, three main aspects were revealed: a) 'grammar and pronunciation', b) 'creativity in presenting information', and c) 'length of the speech and giving adequate details'. Secondly, various differences were found in the level of significance to which the raters attached. Accordingly, the results showed that the university student raters living in the US prioritized the third aspect 'giving adequate details' while the instructor rater group gave importance to the second aspect 'creativity'. The third group opted to focus on the first aspect 'grammar and pronunciation'. The researcher added that these variations could be related to the raters' own dynamics and beliefs regarding teaching and learning speaking skills.

Gui (2012) explored the rating patterns shown by Chinese and American raters working as an EFL teacher in a major university in China. The main data of this study were collected from the scores yielded from a speaking competition, written comments made by the raters, and interviews with the raters. A total number of 6 raters (3 American and 3 Chinese raters working as EFL teachers) participated in the study. The qualitative analyses of this study revealed that while American raters gave importance to the aspects of pronunciation and gestures, Chinese raters remained aloof from these dimensions. Given the difference of rating behaviors regarding pronunciation, the researcher examined this variation in detail by a follow-up interview, resulting in that American raters mentioned specific examples of participants' pronunciation mistakes. However, Chinese raters thought that there were not any pronunciation mistakes affecting the quality of speech. Similarly, American raters were more critical about the 'speech delivery and nonverbal behavior' category than the Chinese raters.

To explore raters' decision-making processes while assessing test-takers' performances in an advanced level speaking test, Pollitt and Murray (1996) conducted a study based on two data collection techniques: a) 'Repertory Grid' and b) 'Method of Paired Comparisons'. A total number of six raters with varying levels of experience contributed to the study. The basic principle of these two methods for the raters was to compare and contrast the performances rather than assigning a score. The data were collected from the speaking section of a high stakes test called the 'Cambridge Certificate of Proficiency in English (CPE)'. As for the tendencies in assessment criteria, this study revealed that the raters relied on the commonalities in pair test takers' performances rather than evaluating the test takers separately. At the same time, in some cases, the raters tended to give rating decisions according to the pair whose proficiency level was lower. In terms of rating approaches, there were two distinct tendencies: synthetic and analytical. Namely, while some of the raters relied on the general view of test takers' performance, others opted to show more neutral and objective rating behaviors by assessing each test taker separately. Furthermore, the study found out that factors such as speech clarity, test takers' characteristic features, and non-verbal behaviors. This naturally raised concerns over the reliability of the assigned scores.

To summarize, scholars in L2 speaking assessment explored rater cognition and raters' decision making processes as regards to three themes. Firstly, the areas that raters give more or less importance have provided a detailed mental picture of raters' decision making processes and the findings of such studies have been a useful contribution to the development of speaking test rubrics. Secondly, rater cognition studies aim to reveal how raters develop and finalize their decisions while evaluating performances. In this regard, whether raters rely on assessment rubric criteria or their own personal beliefs has been examined within the framework of rating approaches. Thirdly, rater characteristics and background factors have been found to have an impact on raters' cognition and decision making patterns. For instance, raters' level of experience and L1 background could be the reasons why raters award lenient or strict scores. Given the effect of raters' background and characteristics on rater cognition, revealing raters' decision making patterns within each specific assessment context could provide solutions to the problems stemming from random rater pair matching and ineffective rater training. The paucity of rater cognition models is another crucial issue in L2 speaking assessment. Therefore, there have been more studies

exploring rater cognition and decision making patterns in L2 writing assessment than L2 speaking assessment. As for these points, this dissertation examined speaking raters' rater cognition and decision making patterns utilizing verbal protocols. Considering that rater cognition and raters' decision making patterns whilst scoring L2 speaking performances in the Turkish context of EFL have not been investigated broadly, this research aimed to fill the gap in the relevant literature.

2.7. Summary and Research Gaps in L2 Speaking Assessment

This chapter consisted of five major areas: the overview of L2 speaking assessment, factors affecting L2 speaking assessment, rater's rating experience, speaking performance and text quality in performance assessment, and rater cognition and decision making patterns while rating L2 speaking performances. The first section highlighted the issues of reliability, validity, and fairness in speaking assessment. The following part in this section expanded on the overview of L2 speaking assessment in the higher education system in Türkiye. As for factors affecting the variations of EFL and ESL speaking scores, the second section reviewed numerous studies by separating into detailed subtopics such as speaking tasks, rater's background, rater training, and rating scales. The next sections provided empirical studies from writing assessment since there were very few studies of speaking assessment research. Therefore, the third section discussed the impact of rater's professional experience on score reliability within the framework of speaking and writing assessment. Similarly, the fourth section provided the findings of studies that examined the effect of text quality on writing score reliability. The last section was very important as it presented writing models of rater cognition by examining implications for transferring them to L2 speaking assessment. In addition, this section gave more details about the issues of rater cognition and decision making patterns while assessing speaking.

As the reviewed studies in the chapter have shown, the impact of rater experience on the variability of speaking scores seems more complicated than it does since the exact definition of rater experience may change depending on the interaction with other factors. For instance, a speaking rater's 10-year experience in conducting IELTS speaking tests might not make him or her experienced in another speaking test. Therefore, it is crucial to

redefine the experience label of raters in the changing contexts. In addition, when the rater experience factor interacts with speaking performance quality, the issue of reliability and fairness may occur. Compared to writing assessment, very little research has been carried out to explore the topics of speaking assessment in the global and Turkish context. This study aims to fill the existing gap in speaking assessment by exploring the effect of rater experience and speaking performance quality on score variation and rater behavior in L2 speaking assessment in a tertiary education context in Türkiye. Given that the issues of rater experience, speaking performance quality and rater cognition have not been researched extensively in speaking assessment, this study elaborates into the varying relationship between rater experience and speaking performance quality as regards to speaking scores and decision making patterns. As for implications, the findings of this study will contribute to the development of institutional speaking assessment models and practices.

As for research methodologies, both quantitative and qualitative methods in L2 speaking assessment studies can be observed. While earlier studies mostly relied on quantitative methods such as classical test theory and item response theory, contemporary studies tended to utilize either qualitative or mixed method research design. G-theory framework and qualitative data collection methods such as verbal protocols and interviews have been effective methods while corroborating the findings retrieved from quantitative methods. Nonetheless, the use of these qualitative methods have not been very popular among speaking assessment scholars due to operational and practical concerns such as transcribing and analyzing the bulk of verbal protocol data. Considering all these points, this research will contribute to the speaking assessment field as regards to both research methodology and statistical framework. At the same time, given the limited number of studies exploring speaking raters' cognition and decision making patterns through verbal protocols, this study aims to fill this research gap in the field.

CHAPTER III

METHODOLOGY

3.1. Introduction

The goal of this dissertation is to explore the effect of rater experience and L2 speaking performance quality on the variation of scores and rater behavior in a Turkish university context. In this study, I utilized convergent parallel case study mixed-method design (see Figure 1). The rationale for using this research design is to gain a greater understanding of a single case, a higher education institution in this study, using quantitative and qualitative data collection tools. In fact, the combination of case study and mixed-method design has the potential for providing insights into the exploration of cases (Yin, 2014).

Although case studies are mostly attributed to qualitative aspects, the inclusion of mixed-method design into case studies would bring certain advantages in terms of interpreting the results and findings (Duff, 2008). Instead of relying on the findings of only one method, employing both quantitative and qualitative data collection tools will provide certain advantages (Mackey & Gass, 2005). Indeed, this research design has been popular in today's research paradigm since it contributes to the development of studies. Despite this fame of mixing methods, it would be crucial to take concerns into consideration while designing mixed-methods research. In other words, the ways in which quantitative and qualitative methods affect each other need to be evaluated well (Strauss & Corbin, 1998). To overcome the complexity of mixing methodologies, a systematic way of organization is regarded as essential (Creswell, 2009; Maxwell, & Loomis, 2003; Tashakkori, & Teddlie, 2003; Yin, 2014). Accordingly, the backbone of this research is based on the main question whether rating experience and varying L2 speaking performance quality have an effect on score variation and rating behavior.

This study adopts the convergent parallel design through which quantitative and qualitative findings were united to approach the results from many different angles. As for the protocol of this design, the quantitative and qualitative data were firstly collected independently. Following the data compilation, overall analyses retrieved from both quantitative and qualitative data were compared and evaluated in the discussion section. Finally, the degree of differences and similarities between databases were interpreted to corroborate the findings (Creswell, 2015; Creswell, & Clark, 2017).

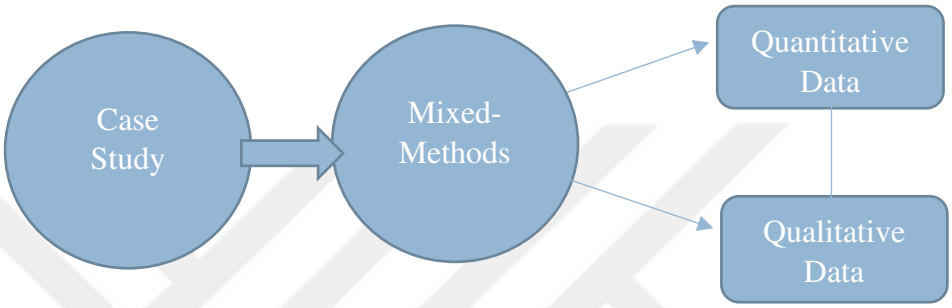

Figure 1. Case study-mixed methods design (adapted from Guetterman, & Mitchell, 2015, p. 2).

Raters contributing to this study were selected from a single EPP to examine the effects of institutional policies of L2 speaking assessment policies on assigned scores and rating behaviors. The L2 speaking performances were collected from the EPP of a state university in Türkiye and utilized as both quantitative and qualitative data.

This chapter begins by examining the theoretical framework and the details of participants who contributed to the study as speaking test raters. In the following sections, data collection instruments and procedures, data analyses, and finally brief information on research ethics are described in detail.

3.2. Statistical Framework

The G-theory approach differs fundamentally from its predecessors, for instance, ANOVA and classical test theory. There are two certain reasons under this difference: G-

theory framework discloses several sources of measurement error and gives researchers useful insights into the details of assessment (Brennan, 2001). In this section, the G-theory framework will be described and explained in detail.

3.2.1. Generalizability Theory as a Statistical Framework

The dichotomy between classical measurement and modern measurement theory has brought scholars to consider the broader context of reliability issues. As such, G-theory and item response theory (IRT) have been used widely in language assessment studies. Although CTT is unsuitable for some certain tests and has serious limitations, it can still be applied in research areas where either G-theory or IRT model is not feasible to utilize (Bachman, 1990). Referring back to the strengths and weaknesses of classical and modern measurement approaches, it would not be wrong to claim that while CTT is able to consider only one source of measurement error at once, G-theory and IRT enable tools that can estimate the degree of various sources of measurement error at the same time (Bachman, 2004). Looking into details of the reliability of assigned scores, I can claim that 'true score' and 'observed score' are two crucial terms to be explained. The real performances of test-takers are directly related to the concept of true score. In addition to this, error score refers to the aspects that are free from the skills to be tested in the assessment design. That is to say, error scores may contain numerous unseen sources of variance. Overall, the mixture of true score and error score explains the concept of observed score (Bachman, 1990; Fulcher, 2010). Given this complexity, CTT might not explain the unsystematic or random source of errors stemming from error score variance (Brennan, 2011).

Considering the major constraints of CTT in performance assessment, G-theory framework enables researchers to work with various source of error at the same time. In fact, G-theory measurement design can provide the concurrent analysis of inter-rater reliability and the casual relationships across single and overall measurement facets. Besides, this statistical framework can improve the reliability of measurement design via various optimization choices (Brennan, 2001; Hsu, 2012). Before a G-study is implemented, the arrangement of research design and the breakdown of analyses need to organized. The initial

step that should be followed in G-theory research design is to give decision whether the study focuses on the dependability and generalizability of the measurement. Secondly, the sources of variance or facets such as raters, items, and contexts need to be selected. Thirdly, the sampling of the facets should be determined. This step is actually related to the degree of estimation of findings. Next, researchers need to decide on whether a crossed or nested design are suitable to the measurement design. After conducting the G-study with a suitable software package, researchers can optimize the generalizability and dependability of coefficient indices during the phase of decision study (D-study). Thanks to this step, they can make necessary improvements in the measurement design (Briesch et al., 2014).

The investigation of source of error variance and score dependability are two mainstays of G-theory measurement design (Shavelson & Webb, 1991). G-studies explore the various source of measurement error as well as the effects these variances on the reliability of given scores. In addition, G-studies assess the impact of variance facets and the object of measurement (students in this study) that contribute to the total variance score (Brennan, 2000, 2001, 2005; Gao & Brennan, 2001). However, D-studies aim to reveal the optimal measurement design within the norm-referenced and criterion-referenced score interpretations. With this in mind, two reliability coefficient indices are salient in D-studies: a) generalizability coefficient (Ep^2) b) dependability coefficient (Φ). While the former one (generalizability coefficient) refers to norm-referenced score interpretation, the dependability coefficient is related to the criterion-referenced situations (Bachman, 1990; Brennan, 2001; Shavelson & Webb, 1991; Suen, 1990).

G-theory analyses can be formed in two types of designs: a) crossed design, and b) nested design. When each student (p) who contributed to a study responded all the speaking performance qualities (q), and all rater experience groups (r) graded all the speaking performances that all the students answer, the facets and object of measurement should be evaluated using a fully crossed design ($p \times r \times q$). However, the reverse scenario requires a fully nested design ($p : r : q$) in which different students answer different quality of speaking performances, and different raters evaluate different speaking performances. In addition, the mixture of crossed and nested designs can be used. For instance, when the first rater assigns the first five speaking performances, the second rater scores the second five speaking

performances, and the third rater evaluates the last five speaking performances. In this scenario, all of the speaking performances are completed by all the students, but different raters score different set of speaking performances [*p x (r : q)*] (Brennan, 2001; Briesch et al., 2014; Taşdelen Teker & Güler, 2019). Considering all these points, G-theory framework including crossed designs all speaking performance qualities (*p x r x q*) and individual speaking performance qualities (*p x r*) was adopted for this study. The sources of variances and their effects on the reliability of assigned scores were investigated. Following this, using the reliability coefficients, ideal measurement scenarios were optimized.

3.3. Selection of Raters

Purposive sampling was utilized in this research. Given the main reason for selecting this sampling type, Mackey and Gass (2005) "researchers knowingly select individuals based on their knowledge of the population and in order to elicit data in which they are interested" (p. 122). Indeed, the researcher had certain knowledge regarding the background of the raters with differing levels of rating experience since both the researcher and the raters were full-time instructors of English at the same EPP. As for the confidentiality of the participants, all raters were presented with pseudo names. A total number of 25 participants contributing to this study were working at a technical university in western Türkiye. Being professionals in foreign language teaching, all the raters were graduates from language related departments such as English Language Teaching (ELT), English Language and Literature (ELL), American Culture and Literature and Linguistics. While a number of 17 raters were Turkish instructors, eight of them were foreign instructors with different L1 backgrounds. A rater profile form (Appendix A), which was adapted from the study (Cumming et. al., 2001), was given to the raters to collect data about their background information such as gender, age, level of education and overall EFL teaching experience both in general and at university levels. Furthermore, the raters were asked to fill out the number of speaking/communication classes they taught in an academic year in their current institution and other higher education institutions. Similar to this, they were requested to fill out the number of speaking assessment duties that they had in their current institution and other higher education institutions. Finally, they were asked to write the number of training

or sessions in speaking assessment and their self-description of themselves as an EFL speaking rater.

The fundamental methodological difference between this dissertation and the relevant studies (Han, 2017; Şahan, 2018) was the categorization of rater experience based on a scale form. Since this dissertation narrowed down its focus on a single institution, it attempted to expand on the existing literature by creating a rater experience scale form (Appendix C). In fact, the definition of rater's in-house rating experience was necessary to reveal the actual speaking rating experience. Otherwise, it would not be possible to differentiate between the year of rating experience and the number of rating experience. Another important consideration about this issue was that experienced speaking raters might have been counted as inexperienced as they had less years of teaching and rating experience. As such, the goal of rater scale form was to reveal the number of rating duties that raters had completed. Given that this dissertation aimed to research the L2 speaking assessment of a single institution, utilizing rater's institutional or in-house rating experience was a major step.

The data that includes the number of classes for teaching speaking, speaking assessment duties and training sessions in speaking assessment were transferred to a rater experience scale form (Appendix C) to categorize raters into low-, medium-, and high experienced rater groups. This rater experience scale form consists of 3 main areas: speaking assessment experience (60%), teaching speaking experience (30%) and training experience in speaking assessment (10%). The total score gathered from these areas determined the raters' experience group. Table 2 illustrates the rater experience groups based on the points from the scale.

Table 2

Rating experience groups of the participants

		Rater experience scale form points (Total)								
		2 pts. or less	*3-4 pts.*	*5-6 pts.*	*7-10 pts.*	*11-15 pts.*	*16-20 pts.*	*21-40 pts.*	*41+ pts.*	**Total**
Experience Group	*Low*	**1**	**2**	**2**	**5**	0	0	0	0	**10**
	Medium	0	0	0	0	**3**	**4**	0	0	**7**
	High	0	0	0	0	0	0	**6**	**2**	**8**
Total		1	2	2	5	3	4	6	2	**25**

As can be seen in Table 2, of the 25 participants, a number of 10 raters that had 10 points or less from the scale were categorized as low-experienced raters. Seven raters having points between 11 and 20 were categorized as medium-experienced raters. The remaining eight raters who had 21 or more points were categorized as high-experienced raters.

Table 3 shows the distribution of raters' gender and age according to the categorization of the experience group.

Table 3

Gender and age distribution of the participants

		Gender		**Age**		
		Male	Female	20-30 years	31-40 years	41-50 Years
Experience Group	Low	2	8	2	7	1
	Medium	2	5	1	4	2
	High	2	6	1	4	3
Total		**6**	**19**	**4**	**15**	**6**

A total number of six male and 19 female raters contributed to the study. While there were two males and eight females in the low-experienced rater group, two males and five females participated in the study as medium-experienced raters. As for the high-experienced raters, there were two males and six females. Given the age distribution of raters, four raters were between 20 and 30 years old while 15 raters were between 31 and 40 years old; six raters were above 40 years old. Among the low experienced raters, there were two raters between 20 and 30 years old while eight raters were over 30 years old. As for the medium-experienced group there was only one rater between 20 and 30 years old; the rest in this group were above 30 years old. Finally, while only one rater was between 20 and 30 years old, four raters were between 31 and 40 years old; three raters were above 40 years old in the high-experienced group. Table 4 gives information on the participants' highest level of education and their background of rater training for speaking assessment.

Table 4

Participants' level of education and previous training on speaking assessment

		Degree			**Previous Training**	
		BA	*MA*	*PhD*	*Yes*	*No*
Experience Group	*Low*	4	5	1	7	3
	Medium	3	4	0	6	1
	High	1	6	1	8	0
Total		**8**	**15**	**2**	**21**	**4**

		Speaking Assessment Training (the number of sessions)				
		2 sessions or less	*3-5 sessions*	*6-8 sessions*	*9-11 sessions*	*12+ duties*
Experience Group	*Low*	8	2	0	0	0
	Medium	3	4	0	0	0
	High	3	3	1	0	1
Total		**14**	**9**	**1**	**0**	**1**

According to the information presented in Table 4, eight of the raters held a BA degree while 15 of the raters completed a MA program. However, only two of the raters held a PhD degree. Considering the raters' previous training background, 21 raters received

training on speaking assessment while four raters had not received any training sessions. Furthermore, while 23 raters received less than six sessions, only two raters participated in over six training sessions.

Table 5

Teaching experience of the participants

		Teaching EFL (total)				
		2 years or less	*3-4 years*	*5-6 years*	*7-10 years*	*10+ Years*
Experience Group	*Low*	0	2	1	3	4
	Medium	0	1	1	3	2
	High	0	0	1	0	7
Total		**0**	**3**	**3**	**6**	**13**
		Teaching EFL in University Context				
		2 years or less	*3-4 years*	*5-6 years*	*7-10 years*	*10+ Years*
Experience Group	*Low*	4	2	1	1	2
	Medium	2	0	2	3	0
	High	0	0	1	3	4
Total		**6**	**2**	**4**	**7**	**6**

Teaching EFL Speaking/Communication at Other Universities (the number of classes)

		5 classes or less	*6-10 classes*	*11-15 classes*	*16-20 classes*	*20+ classes*
Experience Group	*Low*	10	0	0	0	0
	Medium	6	1	0	0	0
	High	5	2	0	0	1
Total		**21**	**3**	**0**	**0**	**1**

Teaching EFL Speaking/Communication in the current institution (the number of classes)

		5 classes or less	*6-10 classes*	*11-15 classes*	*16-20 classes*	*20+ classes*
Experience Group	*Low*	9	1	0	0	0
	Medium	3	4	0	0	0
	High	3	1	2	1	1
Total		**15**	**6**	**2**	**1**	**1**

As can be seen in Table 5, 19 of the raters had over seven years of teaching EFL experience. However, the picture in the raters' teaching EFL experience context was

different and there was less variation than teaching experience in general. As for teaching EFL experience in university context, 13 raters had over seven years of experience while 12 raters had less than seven years of experience in teaching EFL in higher education settings. Given the numbers of speaking classes the raters taught at other universities and in their current institution, 24 of the raters gave less than 11 speaking classes while only one of them instructed more than 20 classes at other universities. Moreover, while 21 raters gave less than 11 classes in their current institution, only four raters had over 11 classes' teaching speaking experience in their current workplace.

Table 6

Assessment experience of the participants

Assessing EFL Speaking at Other Universities (the number of duties)						
		5 duties or less	*6-10 duties*	*11-15 duties*	*16-20 duties*	*20+ duties*
Experience Group	*Low*	9	1	0	0	0
	Medium	5	1	0	1	0
	High	2	1	2	1	2
Total		**16**	**3**	**2**	**2**	**2**
Assessing EFL Speaking in the current institution (the number of duties)						
		5 duties or less	*6-10 duties*	*11-15 duties*	*16-20 duties*	*20+ duties*
Experience Group	*Low*	3	5	2	0	0
	Medium	0	1	2	2	2
	High	0	0	1	2	5
Total		**3**	**6**	**5**	**4**	**7**

According to Table 6, 19 of the raters had less than 11 duties in speaking assessment sessions at other universities while six raters had over 11 assessing speaking duties. When compared with the raters' assessment experience at other universities, it can be said that the variation was totally different with nine raters less than 11 duties and 16 raters over 11 duties in speaking assessment in their current institution.

Furthermore, the raters were asked to rate their self-described speaking assessment experience on a 5-point Likert-type scale with items from "No experience" to "Very experienced." Table 7 gives information on the participants' self-described rating experience according to the experience group determined by the rater experience scale form.

Table 7

Participants' self-described rating experience

		Self-described Rating Experience				
		No experience	*Little experience*	*Some experience*	*Experienced*	*Very experienced*
Experience	*Low*	0	2	4	3	1
Group	*Medium*	0	0	3	4	0
	High	0	0	3	5	0
Total		**0**	**2**	**10**	**12**	**1**

A total number of 12 raters categorized themselves as "experienced" raters, and only one rater described him or herself as "very experienced". Nonetheless, of the 12 raters who described themselves as experienced, only five raters corresponded with the actual high-experienced rater group. While four self-described high-experienced raters were from the actual medium-experienced group, three self-described high-experienced raters were from the actual low-experienced group. Besides, it was striking that one rater from the low-experienced group labelled himself/herself as very experienced. Although this rater reported that he or she scored very little as regards to the number of teaching speaking classes and assessing speaking duties, he or she perceived him or herself as a very experienced rater. Ten raters placed themselves into the category called "some experience." Of the 10 raters with this category, three raters were from the high-experienced group, three raters were from the medium-experienced group, and four raters were from the low-experienced group. Finally, it was quite interesting that only two raters perceived themselves as having "little experience". While any of the raters from the actual medium- and high-experienced groups did not attribute themselves as having "little" or "no experience", only two raters from the

actual low-experienced group described themselves as low-experienced. Based on the raters' self-described rating experience, Table 8 provides information on the self-described experience groups.

Table 8

Participants' self-described rater experience group

		Self-described Rating Experience					
		No experience	*Little experience*	*Some experience*	*Experienced*	*Very experienced*	**Total**
Self-described group	*Low*	0	2	0	0	0	2
	Medium	0	0	10	0	0	10
	High	0	0	0	12	1	13
Total		**0**	**2**	**10**	**12**	**1**	**25**

A total number of two raters who reported that they had either no or little experience were categorized as self-described low-experienced raters; 10 raters that described themselves as having some experience were created as self-described medium-experienced raters; and 13 raters that perceived themselves as experienced or very experienced were labelled as self-described high-experienced raters.

3.4. Data Collection Instruments

Utilizing case study mixed-method design, this study gathered qualitative and quantitative data via a rater profile form, analytic scores assigned to L2 speaking performances, verbal protocols, and written score explanations. The quantitative data consisted of 60 scores given by each rater to L2 speaking performances and the rater profile form providing the background information, through which each rater's experience score was formed. Then, the experience scores were transferred to the rater experience scale form to categorize the raters into experience groups. The qualitative data consisted of verbal

protocols and written score explanations, both of which were used to explore rater decision patterns and behaviors. Besides, the aim of written score explanations that consisted of three reasons was to justify the raters' scores.

3.4.1. Selection of L2 Speaking Performances

The L2 speaking performances were collected from EFL students studying at School of Foreign Languages (SFL) of a technical university in western Türkiye. Mentioning the educational system of this institution, there are four quarters in an academic year, each of which lasts eight weeks. At the end of each academic quarter, various exams such as 'listening-vocabulary-grammar', 'reading' and 'writing' are carried out as well as speaking exams called 'end-of-the quarter final speaking exams'. In addition to these exams, one speaking and writing assignment is given to the students in each quarter. Within the assessment system of this institution, each speaking exam equals to 12 per cent. As for the preparation of speaking test items, a rigorous workflow is conducted with the participation of the testing office, communication/speaking workgroup coordinator and an independent speaking test item committee. Once the preliminary speaking exam questions are prepared by the testing office workgroup members, a hard copy of the speaking test items is sent to the communication workgroup coordinator for a detailed review. Following this, the reviewed items are sent to an independent committee where the members negotiate and evaluate the items to detect any inconsistencies and ambiguities. Then, the edited hard copy of the exam is checked by the testing work group member who is responsible for proofreading and editing. Lastly, the final version of the speaking exam is checked and approved by the testing office coordinator.

During the final week of instruction, students are given a mock exam to make them familiar with the speaking exam format and questions. Before each speaking exam, these mock exam sessions are taken seriously and on test-day conditions are provided for the students. Besides, calibration meetings for speaking assessment are held before each speaking test in an academic year to make sure each instructor is familiar with the institutional speaking exam rubric and speaking assessment policy documents. Thanks to

these sessions, novice or newly recruited teaching staff have the opportunity to adapt the speaking assessment system in this institution. As for test-day conditions, there are certain quality measures such as two examiner systems (one examiner and one rater), examiner guideline forms, a triangle-shaped seating arrangement with the test-taker and the examiners, and optimum physical conditions. Once all the quality steps have been completed, raters accept the students one-by-one by allocating 15 minutes to each student in turn.

Speaking tests are conducted by two examiners. Raters assess each students' performances individually by using separate rubrics. The raters independently fill out the rubrics and assess the performances. Rephrasing of the questions is not allowed and the entire speaking test session is recorded. The recordings are archived in the database of the school. Upon the completion of the exam, the examiners compare their grades and if there is a difference between the assigned scores less than 10 points, the average score is assigned for the related performance. If the difference is more than 10 points, a third rater, preferably the communication workgroup coordinator, assesses the performance utilizing the audio recording. Thanks to double marking and third rater system, the quality loop for the speaking assessment is ensured.

Given the availability of a well-established speaking assessment system in this institution, the preliminary speaking performance data were gathered from the official speaking exam conducted in the academic year of 2017-2018. The proficiency level of the test takers was B1 and labelled as level three in the institution. Before the exam starts, the examiner asks general warm-up questions until the test-taker seems comfortable. The speaking exam consists of three tasks with 10 questions each. In Task 1, the test-taker chooses a numbered card to talk about the topic for no longer than four minutes. In addition, the examiner may ask a follow-up question if the answer is redundant. In Task 2, the test-taker again chooses a numbered card and receives the corresponding question card. Here he/she has one minute to prepare by using a blank sheet of paper and 2 minutes to explain the response. Finally, in task 3, the examiner asks the question corresponding with Task 2. The test-taker should not talk longer than 3 minutes for this task. Due to the confidentiality of the test items, only one sample from each task is presented as in the following:

Task 1

Let's talk about your NEIGHBORHOOD

- In which neighborhood of your town do you live?
- Have you met your neighbors?
- Do you think it is important to have a good relationship with your neighbors?
- How could your neighborhood be improved?

Task 2

Describe the career you hope to have in the future.

You should say:

- What the career is
- What you need to get this career
- What your responsibilities will be in this career

and explain why you are interested in this career.

Task 3

In your opinion, is your career very important for your happiness in life? Please explain your answer.

The availability of a double marking speaking assessment system was already an asset to the quality division process. Considering the categorization of varying quality L2 speaking performances, there were three yardsticks: assigned institutional scores by the double marking assessment system and two independent quality-check raters, one of whom was the communication/speaking workgroup coordinator and the other of whom held a PhD degree in the department of ELT. Both raters were full-time employed instructors of English working in this institution. Besides, they had over seven-year experience in rating L2 speaking performances as well as preparing speaking test items duty. Besides, the communication/speaking workgroup coordinator worked as a third rater in the speaking tests

if there was a disagreement between the scores assigned by two examiners. Assessment instruction form (Appendix D), Analytic Speaking Exam Rubric (Appendix B) and the pack of rubric sheets consisting of speaking exam recordings were delivered to the quality-check raters. The speaking exam rubric was created and developed by this institution for assessing end-quarter speaking exams, placement tests and high stakes tests such as proficiency and student exchange exams.

A total number of 106 L2 speaking performances were collected from the level 3 students for the quality division process. Of the 106 L2 speaking performances, 60 L2 speaking performances, 20 of which were low, 20 of which were medium and 20 of which were high, were selected to be used in this study. Firstly, the quality-check raters assigned scores for each speaking performance using the institutional analytic rubric. After collecting the assigned scores, the researcher compared the scores assigned by the quality-check raters with the ones given by the double examiners. The researcher excluded the L2 speaking performances (n = 21) that had a difference more than 10 points among the scores assigned by the double examiners and quality-check raters. Twenty-five L2 speaking performances were also rejected since some of them had low sound quality and almost 10-point score difference. In fact, the L2 speaking performances that the double examiners and two quality-check raters (4/4 of the quality-check actors) fully agree on low-, medium- and high- quality performances with a very slight score difference were accepted to be utilized in the study. Finally, the researcher determined 20 L2 speaking performances lower than 70 points as low-quality, 20 L2 speaking performances between 70 and 85 points as medium-quality, and 20 L2 speaking performances more than 84 points as high-quality.

Given the importance of quality-division process in this study, a total number of 60 recorded speaking exam performances (20 low-, 20 medium-, and 20 high-quality L2 speaking performances) were chosen out of 106 performances thanks to the double marking system and with the participation of two quality-check raters.

3.4.2. Rating Scale

An analytic rubric with 100-point scoring scale (Appendix B) was utilized in this study. This scoring scale, used as an institutional rubric, was developed by the communication workgroup members. The rubric consists of four main descriptors: task completion/content (...× 8 pts.), vocabulary (...× 5 pts.), grammar and structure (...× 6 pts.), and fluency (...× 6 pts.). These descriptors are multiplied by four categories: unsatisfactory (1 pt.), limited (2 pts.), accomplished (3 pts.), and exemplary (4 pts.). In addition, there are blank areas on the rubric where raters can write their comments as to strengths and weaknesses of the test taker.

The internal consistency of the scale was assessed by composite reliability (CR), and average variance extracted (AVE). Table 9 summarizes the results of CR and AVE scores.

Table 9

The results of CR and AVE for the analytic rubric

L2 Speaking Performance Quality	**CR**	**AVE**
Low-quality speaking performances	.94	.76
Medium-quality speaking performances	.80	.41
High-quality speaking performances	.80	.67

Table 9 shows that the scores of CR for each speaking performance quality ranges from .80 to .94, which should be higher than at least .70. Thus, the scores of CR seemed to show a high degree of internal consistency. As for the AVE scores, it can be seen that they range from .41 to 76, which should preferably be above .50. Even if the AVE score of medium-quality speaking performances is .41, the convergent validity is still sufficient since the CR score is higher than .60 (Fornell, & Larcker, 1981; Lam, 2012). In overall, both CR and AVE scores of the rating scale are within the acceptable threshold scores.

The institution that provided the main data for this study was an internationally accredited EPP. Thus, all procedures regarding testing and assessment quality were recorded and followed by an independent accreditation committee. The institution's consistent excellence in teaching and assessing skills was prioritized by the university administration. As for the requirements for assessment quality, it has been aforementioned that the communication workgroup coordinator organized calibration meetings to make sure raters apply standard and ground rules for the speaking test. Thanks to these sessions, raters had a chance to revise the descriptor labels in the rubric as well as the procedures of the speaking test. In fact, that the rubric was already used as an institutional speaking rubric and regular calibration meetings for speaking exam examiners was an asset to this study as regards to rater orientation. In short, I aimed to minimize the effect of rating rubric on assigned scores utilizing the accreditation quality steps and procedures.

3.4.3. Verbal protocols

It was necessary to carry out a comprehensive training plan since the use of verbal protocols could be challenging for the raters. As such, the researcher formulated a detailed plan to create a mental picture of the process and equip all raters with the principles and use of verbal protocols.

The verbal protocol training process consisted of three main phases: a video-recorded guide, an orientation meeting on the principles of verbal protocol, and one-to-one sessions. Firstly, an expert EFL instructor holding a PhD degree in language assessment and having over seven years of teaching and assessing experience was filmed on a sample verbal protocol process. After having been explained about the purpose of the verbal protocol process and set of instructions, the rater was provided with a voice recorder, a camera set, a laptop, and one sample speaking performance from the preliminary data set. In a single room, the rater graded the L2 speaking performances by following each step of the protocol instruction set. Following this, the researcher uploaded this recorded-video to *YouTube* (Çoban, 2019) and sent it to the raters via email. In the next phase, the researcher organized an orientation meeting to revise the items in the set of instructions as well as highlighting

the points in the sample video. Furthermore, during this meeting, the researcher briefly mentioned the background of verbal protocol use as a data collection instrument in various fields and language assessment studies. After practicing some verbal protocol exercises, the researcher welcomed the raters' questions and promoted an open discussion to clarify the ambiguous aspects in the process. In the last phase, the researcher had one-to-one sessions with the raters to discuss the sample verbal protocol video in detail and reflect on the open discussions during the orientation meeting.

3.4.4. Written Score Explanations

The raters were asked to explain the reasons for their assigned scores. While doing so, they were requested to classify their score explanations according to either positive or negative connotations. The data retrieved from the score explanations were used to corroborate the findings from the verbal protocols.

3.5. Data Collection Procedures

Following the quality division of the data and the orientation process for verbal protocols, the raters were provided with a data pack consisting of a USB with recorded L2 speaking performances, analytic scoring rubrics, a background questionnaire, and a set of instructions for verbal protocol. The raters were informed to finalize the rating process in a three-month period starting from June to September 2019. Given that there was less workload in the summer time period for the raters in this institution, the researcher chose this time period so that the raters could focus on their rating duty efficiently.

Quantitative and qualitative data collection methods were utilized in this study. A total number of 7,500 scores (1,500 total scores and 6,000 sub-scores) were comprised of quantitative data while 375 verbal protocols and 4,500 written score explanations were made up of the qualitative data. Thanks to the verbal protocol instruction video, clear set of instructions, one-to-one orientation sessions, and the raters' familiarity with the institutional

speaking exams, all raters successfully finalized the rating of the L2 speaking performances and verbal protocols. In addition, the researcher contacted each rater in person on a regular basis to track the progress of verbal protocol process. As mentioned in the set of instructions, the instructors were free to opt for the language (either English or Turkish) that they would use while carrying out verbal protocols.

3.5.1. Rating procedure

Utilizing a 100-point analytic scoring rubric, the raters gave their scores according to the four categories (unsatisfactory, limited, accomplished, and exemplary) and four descriptors (task completion/content, vocabulary, grammar & structure, and fluency). The L2 speaking performances (low-, mid-, and high-quality) were arranged randomly so that the raters would not be affected by the order of the data set. In case the raters need, they were reminded that it was possible to assign partial points based on the boundaries of the descriptors in the rubric. Given the importance of standardizing the process of rating, a set of instructions for verbal protocols (Appendix E), through which the basics of the research and the principles of assessing the performances, were provided to the raters. In this set of instructions, it was underlined that the raters had better evaluate each speaking performance by itself and would not compare one another. Another crucial point was that the raters were not allowed to negotiate their given scores with other parties. Additionally, the raters were requested to fill out the written score explanations in the space given in the rubric. The rationale behind this was to reveal the patterns that the raters followed while grading the L2 speaking performances.

3.5.2. Recording raters' thoughts

During the rater orientation meeting, the raters were instructed how to record their voices while assessing the L2 speaking performances. The raters were reminded that retrospective verbal report technique need to be taken into consideration primarily since the recording would be playing while assessing the L2 speaking performances. However, they were also reminded they were free to utter simultaneously what they were thinking about the

rating. Most importantly, either retrospective verbal reports or simultaneous think-aloud protocols, the ultimate goal was to make raters say everything they thought and recorded safely onto the voice-recorder. The researcher marked each of 15 L2 speaking performances by mentioning the codes in the file names (e.g. VP SP001, VP SP019, and VP SP053) to prevent any confusion.

As the raters' identities were based on confidentiality, the raters were reminded that even if the things that they uttered were trivial, they needed to keep talking and report fully. Given the reliability of the data, the raters were to complete each attempt at once. Thus, once they took a break while assessing the performances, they had to indicate on the recording that they ended and started again the rating process. Another important point was that the raters were to carry out the assessment process as naturally and as honestly as they could. The researcher underlined that the raters were not supposed to rationalize their ideas at length, and however, they were to reveal their natural thought process while making the decisions. As for the language choice, the raters were allowed to opt for either Turkish or English. Upon the completion of the rating process, the participants delivered the packs to the researcher in person.

3.6. Data Preparation

Qualitative and quantitative methods were utilized in this study. Microsoft Excel was the main data preparation hub for both qualitative and quantitative data. As for data analysis, various software packages were used. SPSS Statistics 25.0 was used for descriptive and inferential statistics while analyzing the quantitative and qualitative data while EduG 6.0 program was opted for generalizability analysis for quantitative data (Briesch et al., 2014). Additionally, NVivo 12 Pro and Amazon AWS Transcribe software packages were utilized for transcribing and analyzing the qualitative data set.

3.6.1. Preparing the quantitative data

The L2 speaking performances that raters gave scores formed the quantitative data. Microsoft Excel was utilized to record 7,500 scores (1,500 total scores and 6,000 sub-scores) into the layout. Thanks to this system, it was possible to detect any discrepancies in the total scores that the raters had summed up.

3.6.2. Descriptive and Inferential Statistics

Using SPSS, descriptive and inferential statistics were employed to reveal whether there were any significant differences among the scores that varying rater experience groups assigned to the low-, mid-, and high-quality L2 speaking performances. While carrying out these statistics, both total scores and sub-scores of the L2 speaking performances were utilized. In addition to this, descriptive statistics were used for the qualitative data: the codes from the VPA and written score explanations.

3.6.3. G-theory Analysis

EduG 6.0 software was used to estimate the sources of score variation that contribute most relatively to the score variability in the analytic scores of L2 speaking performances. As for the investigation of score reliability within speaking performance qualities and rater experience groups, both generalizability and dependability coefficient indices were obtained. In this study, the object of measurement refers to students that were presented as persons. All the participating raters were shown as raters and low-, medium-, and high-quality L2 speaking performances were presented as qualities. In essence, a fully crossed design, namely person-by-rater-by quality (*p x r x q*) was adopted in this study. Person-by-rater-by quality (*p x r x q*) was computed for all mixed quality L2 speaking performances. Besides, person-by-rater (*p x r*) was carried out for each speaking performance quality and rater experience group.

3.6.4. Preparing Qualitative Data

Verbal protocols collected from 15 L2 speaking performances and written score explanations with three reasons from 60 L2 speaking performances, both of which were used to corroborate the data yielded from the speaking performance scores, were the main qualitative data. In fact, verbal protocols and three reasons that raters provided aimed at triangulating and therefore establishing the reliability and validity of the research methodology of this study.

3.6.5. Transcribing and Coding Verbal Protocols

Qualitative content analysis was employed while analyzing the data yielded from the verbal protocols. To achieve this, an inductive approach, namely bottom-up process, was adopted. Starting with the first step, the researcher transcribed the verbal protocols collected from the raters that contributed to the study. The length of the verbal protocols was 51 hours and 21 minutes. While a total of 17 hours and 3-minute length verbal protocols were recorded by high-experienced raters, 13 hours and 51-minute length verbal protocols were recorded by medium-experienced raters; 10 hours and 27-minute length verbal protocols were recorded by low-experienced raters. Considering the heavy workload of the transcription process, Amazon AWS Transcribe were utilized to include speech to text features in the study. Unlike outdated speech to text software or applications, Amazon AWS adopts a deep learning process, by means of which automatic speech recognition feature gave accurate and clear results specifically with L2 speaking performances in English. Despite the high level of accuracy in transcriptions of English, the transcriptions in Turkish required more corrections. Therefore, the researcher edited each automatically transcribed English and Turkish texts by listening carefully to make necessary changes and corrections. Secondly, the researcher determined the principles of data segmentation to divide the coding parts into independent units. Adapting the three criteria set by Cumming et al. (2002, p. 76): “a) by pauses of 5 seconds or more, b) by the rater reading aloud a segment of the composition, or c) by the start or end of the assessment of a single composition”, the researcher established three criteria: a) by pauses of 5 seconds or more (/), b) once the rater utters a new and

meaningful coding unit, and c) when the rater begins or finalizes the assessment of an independent speaking performance. In the transcribed texts, the pauses were separated with a slash mark (/).

Following the data segmentation process, a comprehensive inductive content analysis was planned and executed. According to Mackey and Gass (2005), inductive data analysis necessitates the process of utilizing any meaningful and relevant pieces of raw data to reach major findings. With this in mind, there were two reasons for which the researcher adopted this type of analysis: a) a coding scheme for a L2 speaking assessment design was not available in the relevant literature, and b) most of the information as to verbal protocol coding was either fragmented or based on L2 writing assessment research. Initially, the researcher read and examined each of the transcribed spoken test chosen for the piloting stage. This preparation phase by reading and reading in detail enabled the researcher to be familiar with the transcribed speaking performance texts. Later on, the researcher selected three transcribed texts randomly from each rater in order to create a data set for this stage. Staying in line with the research questions referring to the qualitative data, the researcher coded a total number of 75 transcribed L2 speaking performances thanks to NVivo 12 Pro. While coding the selected transcribed texts, the researcher followed three fundamental phases: a) data reduction, b) data grouping, and c) the formation of concepts. In other words, the researcher analytically questioned the raw data by combining the relationships between sub-themes and main themes (Kyngäs, 2020).

Giving the details of the inductive approach, the researcher firstly skimmed and scanned each sentence of the raw data to exclude the irrelevant themes from the data. For example, the researcher omitted the parts where raters mentioned the procedures of the speaking test, the reminders of the test duration during the exam, and some meaningless statements. At the same time, by determining relevant units of raw data, namely, open codes, the researcher carried out the ‘data reduction’ stage. Next, the researcher analyzed these open codes by comparing and contrasting the similarities and differences so as to form sub-themes and main themes. The researcher displayed open codes such as ‘very little hesitation’, ‘almost no hesitation or fillers’, ‘clear pronunciation’, ‘good comprehension of questions’, ‘adequate and mostly elaborated answers for the task’, ‘a few grammar mistakes’, ‘word

order not so bad', and 'a few wrong word choices'. Following that, the researcher associated the open codes with various sub-themes such as assessing fluency, assessing grammar, and assessing lexical knowledge. This stage can be called data abstraction. Finally, by grouping the sub-themes, the researcher formed the main themes such as language-related strategies and content-related strategies (Dörnyei, 2007; Kyngäs, 2020). After carrying out the coding process, the researcher formed a preliminary coding scheme. However, the major categories and some of the sub-themes of this scheme was adapted from Cumming et al. (2002).

Utilizing the preliminary coding scheme, an independent researcher, who held a PhD degree in language assessment and had expertise in rating performance assessment, coded a total number 75 transcribed L2 speaking performances that the researcher had randomly selected. This piloting data set for interrater reliability corresponded to 15 percent of the total verbal reports. Cohen's κ analysis between the researcher and independent researcher was found to be $\kappa = .92$ with $p < .001$, which suggested a very good agreement between the coders (Cohen, 1960; Landis & Koch, 1977).

3.6.6. Thematic Content Analysis for Written Score Explanations

A total number of 4,500 written score explanations were obtained through all speaking performance ratings by all rater experience groups. Thematic content analysis was utilized while analyzing the explanations made by the raters. The coding pattern in which both focus and type of the explanations were determined was adapted from the study conducted by Barkaoui (2010b). As for focus, themes such as fluency, pronunciation, and grammar use were coded. Additionally, positive and negative connotations that raters attributed to explanations were investigated. However, the raters did not provide neutral explanations as the researcher had informed that they would only use either positive or negative comments. The rationale behind was to elicit clear explanations regarding the description of performances. Also, an interrater reliability analysis was conducted with the data including 10 percent of the total score explanations. The analysis was performed to determine high level of consistency between the two coders, $\kappa = .85$ with $p < .001$.

3.7. Ethical Considerations

The principles of research ethics were taken into account in the entire process of this dissertation. First, all students from whom the L2 speaking performances were collected were informed by a consent form that there were not any risks or harm arising from research. Second, all the participating raters and the expert raters were also asked for their voluntarily contribution to the study. Additionally, all participants (raters and students) were informed about the research methodology of the study and were reminded that they could withdraw from the research at any time. It was underlined that all the personal data and the identities of participants are stored securely and confidentially, and can only be used for intended purposes. Overall, the researcher obtained the official ethics approval from the Ethics Committee of Graduate School of Social Sciences and Educational Sciences at Çanakkale Onsekiz Mart University.

CHAPTER IV

RESULTS

4.1. Introduction

This chapter provides information on the results of data analysis. As for research design, case study mixed-method design was utilized in this study. While four of the research questions refer to quantitative results, the latter two of them are about qualitative results. Each data analysis result has been presented under the related research question separately. Initially, the chapter provides brief information on the characteristics of sample, then details the findings of research questions based on quantitative data analysis. Finally, the chapter elaborates the results of qualitative data analysis.

4.2. Sample Characteristics

Various normality test components such as Kolmogorov-Smirnov, and Shapiro-Wilk as well as the visual interpretation of histograms, normal Q-Q plots, and box plots were examined for each speaking performance data set. Table 10 provides information on the results of Kolmogorov-Smirnov, and Shapiro-Wilk tests.

Table 10

Test of normality results for L2 speaking performance quality groups

	Kolmogorov-Smirnov[a]			**Shapiro-Wilk**		
	Statistic	df	*p*.	Statistic	df	*p*.
Low-quality L2 speaking performances	.051	500	.003	.995	500	.080
Medium-quality L2 speaking performances	.065	500	.000	.995	500	.110
High-quality L2 speaking performances	.059	500	.000	.984	500	.000

a. Lilliefors Significance Correction

As can be seen in Table 10, all the *p*-values of both Kolmogorov-Smirnov were smaller than .05, which illustrated that the data might not be normally distributed. While the Shapiro-Wilk test showed that the *p*-values of low-quality and medium-quality speaking performance data sets were greater than .05, it suggested an opposite finding for the high-quality speaking performance data set. All in all, the normality tests and a visual examination of their normal Q-Q plots, histograms, and box plots showed that the total of speaking performance scores were not approximately normally distributed with a skewness of -0.100 (SE = 0.11) and a kurtosis of 0.297 (SE = 0.22) for the low-quality data set, a skewness of -0.026 (SE = 0.11) and a kurtosis of -0.152 (SE = 0.22) for the medium-quality data set, and a skewness of -0.168 (SE = 0.11) and a kurtosis of -0.609 (SE = 0.22) for the high-quality data set (Doane & Seward, 2011; Razali & Wah, 2011).

4.3. Quantitative Data Analysis Results

As regards to exploring the first two research questions (RQ), SPSS 25.0 was utilized while conducting descriptive and inferential statistics. The aim of **RQ1** was *whether there were any significant differences among the analytic scores of low-, medium- and high-quality L2 speaking performances*. However, **RQ2** revealed *whether there were any significant differences among the analytic scores assigned by low-, medium- and high*

experienced raters. That is to say, while speaking performance quality was the main focus of the first research question, varying rater experience was for the second one. G-theory framework was used for **RQ3** and **RQ4,** the former of which investigated *the sources of score variation that contribute most (relatively) to the score variability of the analytic scores of L2 speaking performances* and the latter of which examined *whether the reliability of the analytic scores of raters (low, medium and high) differ from each other.*

4.3.1. Results for RQ1

RQ1: Are there any significant differences among the analytic scores of low-, medium- and high- quality L2 speaking performances?

Descriptive and inferential statistics were carried out to reveal the findings and the results are presented in figures and tables.

Figures 2, 3 and 4 show the discrepancies of median values and the range of scores assigned to high-, medium-, and low-quality L2 speaking performances. Presenting the dispersion of data can be regarded as one of the strengths of boxplots since they enable us to see the data into quartiles. As for the structure of boxplots, there are five main areas: the minimum score, lower quartile, median, third quartile and the maximum score. Namely, from minimum score to maximum score, each section of the boxplot corresponds to 25% of the data set. In other words, top, middle and bottom fall to 25%, 50% and 25% of the data respectively. Median values are presented by a horizontal bar inside the box. There are two whiskers that stick out from the top and bottom of the boxplot, which shows the lowest and highest values. In addition to this, boxplots show outliers with little circles including numbers. While the first figure (n=20) gives information on the scores assigned to high-quality L2 speaking performances, the second (n=20) and the third boxplots (n=20) show the distribution of the scores assigned to medium-, and low-quality L2 speaking performances respectively.

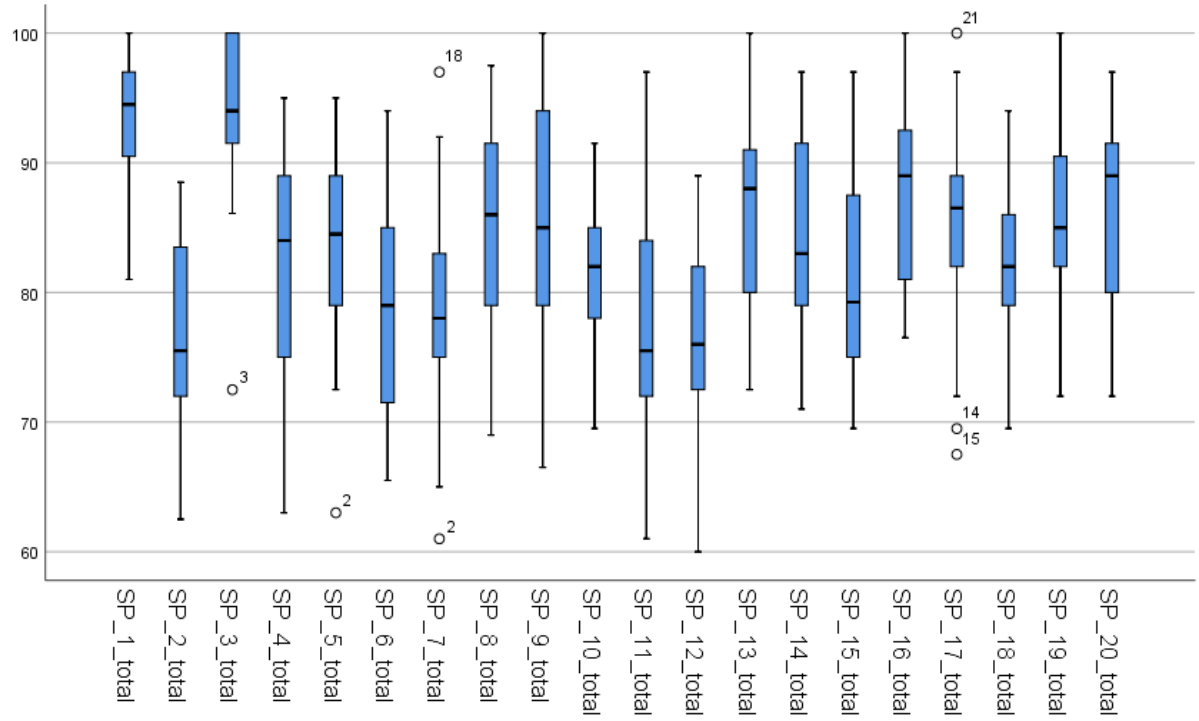

Figure 2. Boxplots for the total scores assigned to high-quality L2 speaking performances

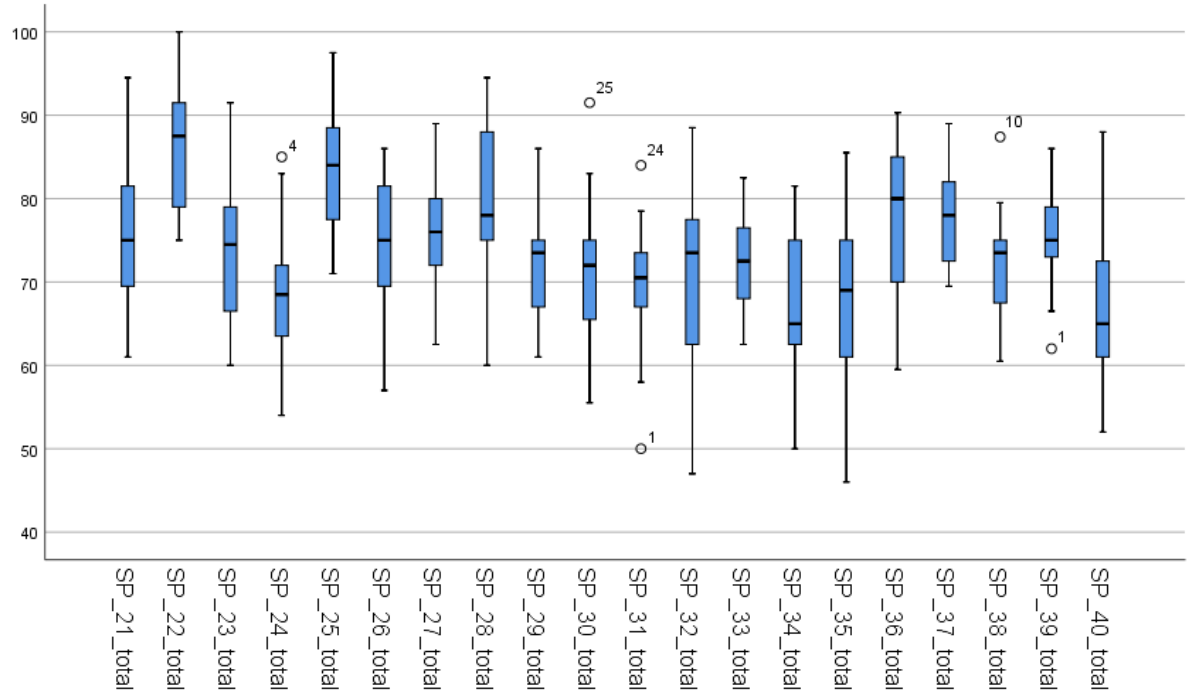

Figure 3. Boxplots for the total scores assigned to medium-quality L2 speaking performances

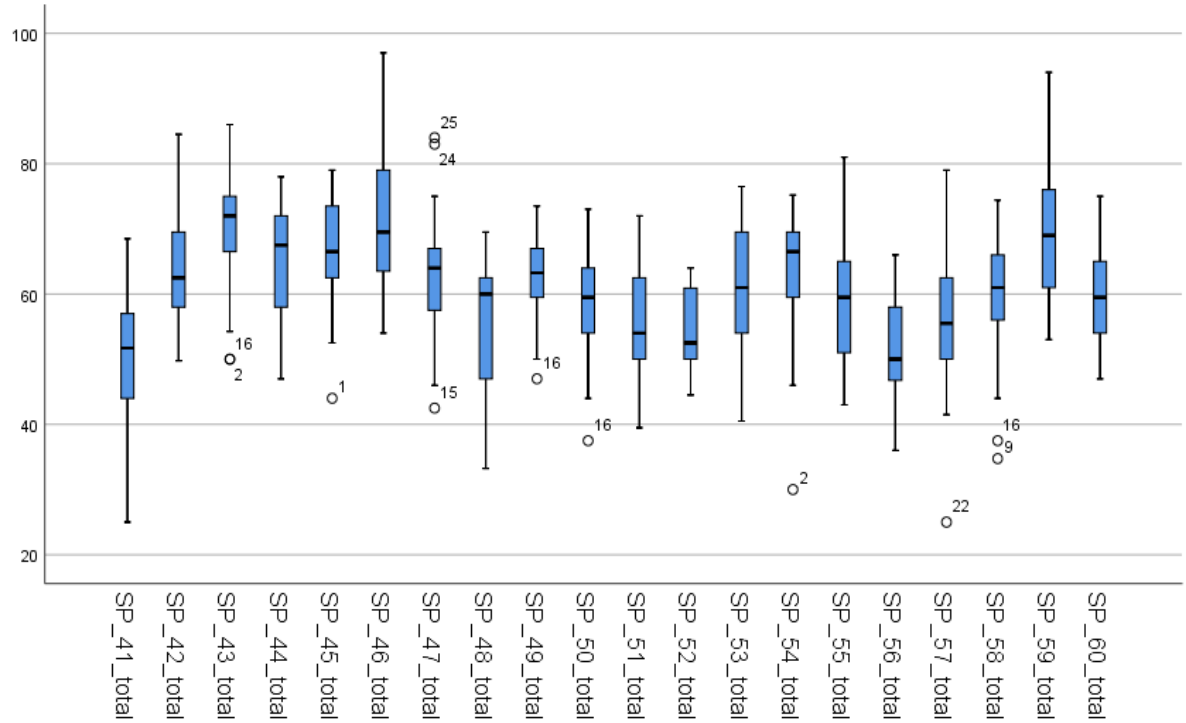

Figure 4. Boxplots for the total scores assigned to low-quality L2 speaking performances

By examining the boxplots, we can see that the dispersion of scores is larger for high-quality L2 speaking performances. While half of the median scores are closer to the upper quartile, the rest of them stay between the minimum score and the first quartile. In addition to this, the length of whiskers shows a large variation both in the lower and higher end of the boxplots. As regards to medium- and low-quality L2 speaking performances, we can also visualize larger ranges in the quartiles although the minimum and maximum scores display a relatively smaller variation than high-quality speaking performance scores. Indeed, the median scores of medium- and low-quality L2 speaking performances seem to display more variance.

Although the boxplots helped to summarize the deviation of the median scores across three varying quality L2 speaking performances, I relied on descriptive and inferential statistics to simply present the data and make inferences about the scores of L2 speaking performances. Computing the range between minimum and maximum scores for each speaking performance (see Appendix G for high-quality L2 speaking performances, Appendix H for medium-quality L2 speaking performances, and Appendix I for low-quality

L2 speaking performances), the mean range for all L2 speaking performances was 32 and the mean range for high-quality L2 speaking performances was 28 while the mean range for mid-quality and low-quality L2 speaking performances were 30 and 36 respectively.

Following this, a Kruskall-Wallis test was conducted to reveal whether there are any significant differences among the analytic scores of varying quality L2 speaking performances. Table 11 summarizes the findings of the test results across low-quality, medium-quality, and high-quality L2 speaking performances.

Table 11

Kruskall-Wallis test results for speaking performance quality groups

H **(2, *n* = 60)**	***p***	**Low-quality speaking performances *(Mdn)***	**Medium-quality speaking performances *(Mdn)***	**High-quality speaking performances *(Mdn)***
46.32	.00	60.00	73.56	83.57

The test revealed a statistically significant difference in the analytic scores assigned to low-, medium- and high-quality L2 speaking performances [(Gr1, n high-quality L2 speaking performances = 20; Gr2, n medium-quality L2 speaking performances = 20; Gr3, n low-quality L2 speaking performances = 20), H (2, n = 60) = 46.32, p = .00)]. The high-quality L2 speaking performances were given a higher median score (*Mdn* = 83.57) than the medium-quality (*Mdn* = 73.56) and the low-quality L2 speaking performances (*Mdn* = 60.00).

After running the Kruskall-Wallis test revealed significant differences across three different quality L2 speaking performances, I used a Mann-Whitney *U* test to compare all pairs of groups whether they were statistically significant from each other.

Table 12

Mann-Whitney *U* test results for speaking performance quality groups

Quality Groups	*n*	*Mdn*	*U*	*z*	*p*	*r*
Low	20	60.00				
			15.00	-5.00	.00	.79
Medium	20	73.56				
Low	20	60.00				
			0.00	-5.41	.00	.85
High	20	83.57				
Medium	20	73.56				
			34.00	-4.49	.00	.71
High	20	83.57				

The test displayed statistically significant differences between low-quality (*Mdn* = 60.00, *n* =20) and medium-quality [(*Mdn* = 73.56, *n* =20) speaking performance groups, *U* = 15.00, *z* = -5.00, *p* = .00, *r* = .79]. Similarly, the test indicated that the difference was statistically significant between low-quality (*Mdn* = 60.00) and high-quality [(*Mdn* = 83.57) speaking performance groups, *U* = .00, *z* = -5.41, *p* = .00, *r* = .85]. Finally, the test revealed statistically significant results between medium-quality [(*Mdn* = 73.56) and high-quality (*Mdn* = 83.57) speaking performance groups, *U* = 34.00, *z* = -4.49, *p* = .00, *r* = .71]. The Mann-Whitney *U* test results displayed that there are significant differences among the analytic scores assigned to high-, medium-, and low-quality L2 speaking performances.

4.3.2. Results for RQ2

RQ2: Are there any significant differences among the analytic scores assigned by low-, medium- and high experienced raters?

As explained in detail in the methodology chapter, the raters were placed in their experience group based on the score retrieved from rater experience scale form. Assessment experience (60%), teaching speaking experience (30%) and speaking assessment training experience (10%) are the main sections in this form. Raters with 20 to higher experience points were categorized in the high-experienced group (n = 8); raters with 19 to 10 experience points were labelled as medium-experienced group (n = 7); raters with 9 to lower experience points were classified as low-experienced group (n = 10). Figure 5 gives information on the mean scores for each of the high-quality L2 speaking performances by rater experience groups.

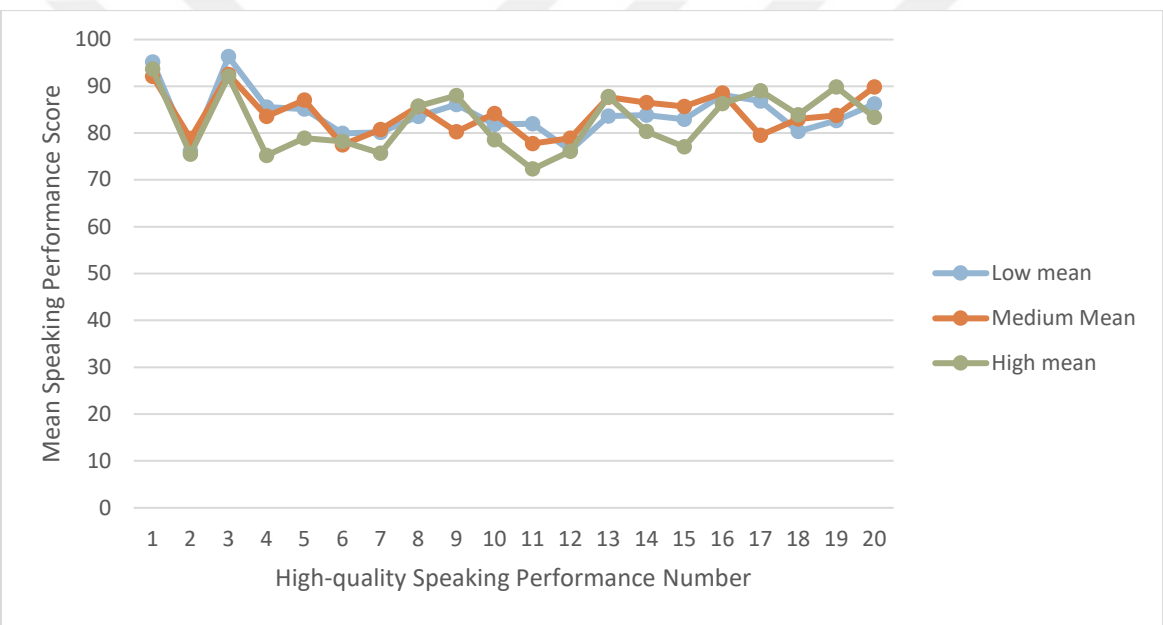

Figure 5. Scores assigned to high-quality L2 speaking performances according to rater experience

As can be seen from Figure 5, raters in all three groups showed a similar pattern although the high-experienced raters seemed to give relatively lower scores to some of the responses such as speaking performance 4 and 5. Despite these slight differences in the given scores, low-, medium- and high-experienced raters tended to give similar mean scores (84.13, 84.16 and 82.38) to high-quality L2 speaking performances.

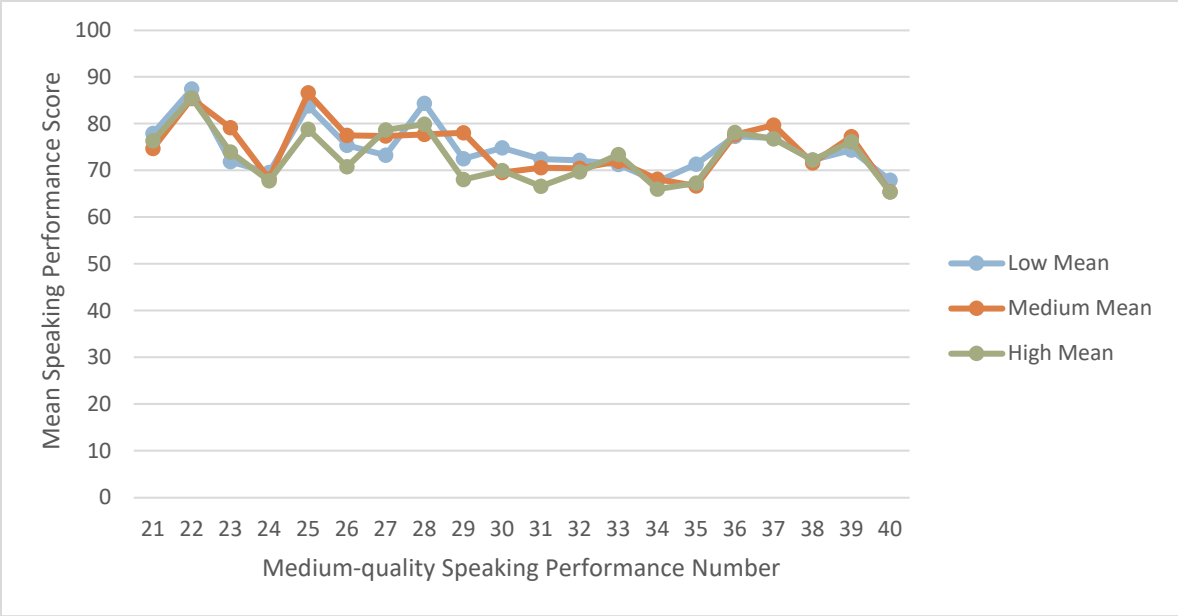

Figure 6. Scores assigned to medium-quality L2 speaking performances according to rater experience

According to the information in Figure 6, the same overall trend can be observed with slight differences among three experience groups. Performances below 70 points were determined as low-quality L2 speaking performances. As such, the expectation from all rater groups was to assign scores higher than 70 points to the medium-quality responses. When the data in the figure is examined in detail, it can be seen that all rater groups tended to give less than 70 points to some of the performances such as number 24, 34 and 40. However, the mean scores given to medium-quality L2 speaking performances by low-, medium-, and high-experienced raters showed similarities (74.72, 74.66, and 73.06 respectively).

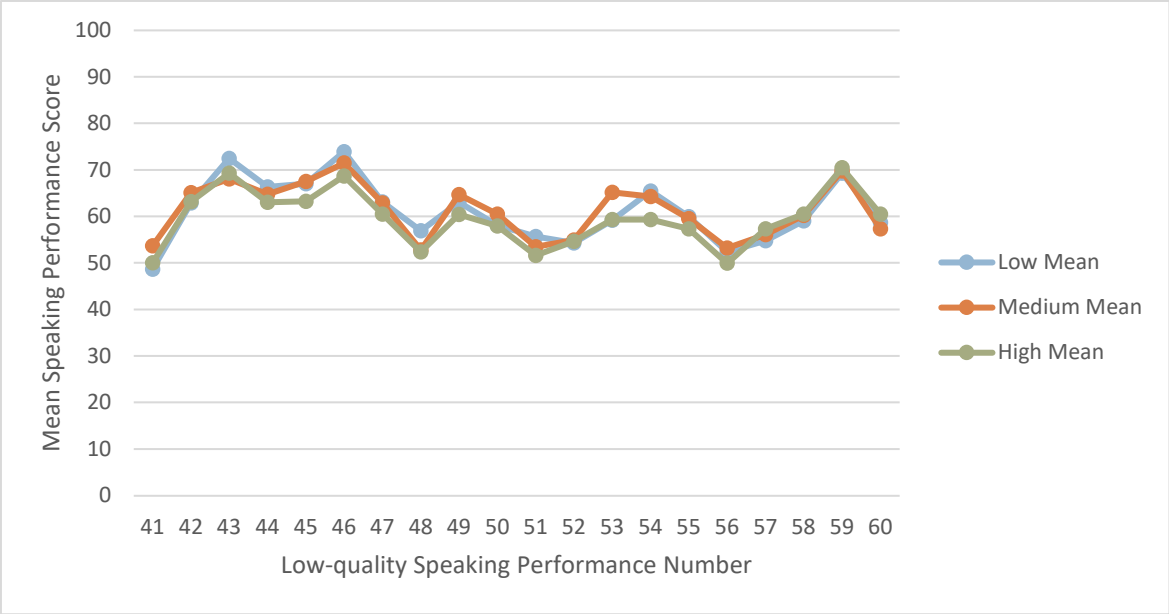

Figure 7. Scores assigned to low-quality L2 speaking performances according to rater experience

As evident from Figure 7, it can be observed that all rater groups showed a similar pattern with a few slight fluctuations. Given the mean scores assigned to low-quality L2 speaking performances, low-, medium-, and high-experienced raters gave similar scores (61.04, 61.25, and 59.47, respectively).

Table 13

Mean speaking performance scores by experience groups

		Mean Score based on L2 speaking performances		
		Low-quality	Medium-quality	High-quality
Experience	Low	61.04	74.72	84.13
	Medium	61.25	74.66	84.16
	High	59.47	73.06	82.38

As can be observed in Table 13, low-experienced and medium-experienced raters tended to assign higher scores to low-, medium-, and high-quality L2 speaking performances while high-experienced raters tended to give slightly lower scores to three different types of L2 speaking performances.

Having examined the overall tendencies, three rater experience groups (Gr1, $n_{\text{low-experienced raters}} = 10$; Gr2, $n_{\text{medium-experienced raters}} = 7$; Gr3, $n_{\text{high-experienced raters}} = 8$) were compared using non-parametric tests. While doing so, Kruskall-Wallis tests were carried out to reveal whether these overall tendencies were statistically significant. Table 14 shows the Kruskall-Wallis test results for low-quality L2 speaking performances across three rater groups.

Table 14

Kruskall-Wallis test results for low-quality L2 speaking performances across rater experience groups

H **(2, *n* = 25)**	***p***	**Low-experienced raters[a]** ***(Mdn)***	**Medium- experienced raters[b]** ***(Mdn)***	**High- experienced raters[c]** ***(Mdn)***
.53	.76	62.10	64.93	58.82

$n^a = 10$ raters. $n^b = 7$ raters. $n^c = 8$ raters

A Kruskall-Wallis test revealed no statistically significant differences in the mean scores assigned to low-quality L2 speaking performances across rater experience groups [(Gr1, $n_{\text{low-experienced raters}} = 10$; Gr2, $n_{\text{medium-experienced raters}} = 7$; Gr3, $n_{\text{high-experienced raters}} = 8$), $H(2, n = 25) = .53$, $p > .05$]. The medium-experienced raters assigned a higher median score ($Mdn = 64.93$) than the low-experienced raters ($Mdn = 62.10$), and high-experienced raters ($Mdn = 58.82$). In addition to this, Kruskall-Wallis tests were computed for each low-quality speaking performance score as well as their rubric component scores across three rater groups. The results showed that there were no statistically significant differences in each total score and component score given to low-quality L2 speaking performances by rater groups.

Table 15

Kruskall-Wallis test results for medium-quality L2 speaking performances across rater experience groups

H **(2, *n* = 25)**	***p***	**Low-experienced raters[a]** ***(Mdn)***	**Medium-experienced raters[b]** ***(Mdn)***	**High-experienced raters[c]** ***(Mdn)***
1.11	.57	73.78	75.43	73.25

n^a = 10 raters. n^b = 7 raters. n^c = 8 raters

There were no statistically significant differences in the mean scores assigned to medium-quality L2 speaking performances across the three rater experience groups [(Gr1, $n_{\text{low-experienced raters}}$ = 10; Gr2, $n_{\text{medium-experienced raters}}$ = 7; Gr3, $n_{\text{high-experienced raters}}$ = 8), H (2, n = 25) = 1.11, $p > .05$]. Similar to the scores assigned to low-quality L2 speaking performances, the medium-experienced raters assigned a higher median score (*Mdn* = 75.43) than the low-experienced raters (*Mdn* = 73.78), and the high-experienced raters (*Mdn* = 73.25), both of which were quite similar.

After the Kruskall-Wallis test results that provided non-significant differences in the total scores of medium-quality L2 speaking performances assigned by three rater experience groups, Kruskall-Wallis tests were conducted on each medium-quality speaking performance score as well as their rubric component scores. There were not any statistically significant differences for the rubric component scores assigned by the three rater groups for medium-quality L2 speaking performances. However, the findings illustrated there was only one statistically significant differences in the scores assigned to speaking performance 29, [H (2, n = 25) = 8.15, p = .017]. For this speaking performance, Mann-Whitney U tests were conducted to determine the statistically significant pairs among rater experience groups. The tests revealed that there was a statistically significant difference between the medium-experienced (*Mdn* = 78.4, n = 7) and high-experienced raters [(*Mdn* = 67.5, n = 8), U = 3.50, z = -2.84, p = .002, r = .73]. No statistically significant results were revealed between low-, and medium-experienced; low-, and high experienced raters.

Table 16

Kruskall-Wallis test results for high-quality L2 speaking performances across rater experience groups

H **(2, *n* = 25)**	***p***	**Low-experienced raters[a]** ***(Mdn)***	**Medium-experienced raters[b]** ***(Mdn)***	**High-experienced raters[c]** ***(Mdn)***
2.22	.33	83.46	84.25	82.18

n^a = 10 raters. n^b = 7 raters. n^c = 8 raters

No statistically significant differences were found in the mean scores assigned to high-quality L2 speaking performances across the three rater experience groups [(Gr1, $n_{\text{low-experienced raters}}$ = 10; Gr2, $n_{\text{medium-experienced raters}}$ = 7; Gr3, $n_{\text{high-experienced raters}}$ = 8), H (2, n = 25) = 2.22, $p > .05$] even if the medium-experienced raters tended to give a higher median score (*Mdn* = 84.25) than the low-experienced raters (*Mdn* = 83.46), and the high-experienced raters (*Mdn* = 82.18). Furthermore, Kruskall-Wallis tests were conducted on each high-quality speaking performance score and their rubric component scores across three rater groups. The results illustrated that there were no statistically significant differences in each total score and component score assigned to high-quality L2 speaking performances by low-, medium-, and high-experienced raters.

4.3.3. Results for RQ3

RQ3: What are the sources of score variation that contribute most (relatively) to the score variability of the analytic scores of L2 speaking performances?

The person-by-rater-by-quality (p *x* r *x* q) G-study design was implemented to examine the sources of score variation that have an effect on the analytic scores that raters awarded. Table 17 provides information on the analysis of score variation stemming from the various sources of variance.

Table 17

Analysis of variance for random effects *P X R X Q* design

Source of Variance	*df*	σ^2	%
P	19	96.79	49.9
R	24	-2.91	0
Q	2	-0.76	0
PR	456	24.59	12.7
PQ	38	0.56	0
RQ	48	18.01	9.3
PRQ	912	54.51	28.1
Total	1499		100

According to the findings retrieved from the G-theory analysis provided in Table 17, persons (P) was the largest source of variance with 49.9%, showing that students differed in their L2 speaking performances. Additionally, this result was expected because the goal of assessment tasks is to distinguish the speaking abilities of students. The residual (PRQ), which actually referred to the interaction of speaking performance quality, rater experience groups, and other unknown source of variances, was the second largest source of variance with 28.1%. The interplay between person and raters (PR) was the third biggest variance that had an effect on score variability with 12.7%, indicating that certain raters showed inconsistencies while grading speaking performances. The next largest variance was between raters and speaking performance quality (RQ) with 9.3%, illustrating that some of the raters showed variance while assessing speaking performances of different qualities. Finally, the other variance sources such as rater (R), quality (Q), and the interaction between persons and speaking performance quality (PQ) did not make any contribution to the variability of the assigned scores.

The person-by-rater (p *x* r) random effects G-study design was conducted to reveal the sources of variation that contribute to the scores awarded to low-quality L2 speaking performances. Table 18 summarizes the sources of variance affecting the score variability of low-quality L2 speaking performances, and compares the interaction between these variance components.

Table 18

Analysis of variance for random effects p x r design (low-quality L2 speaking performances)

Source of Variance	*df*	σ^2	%
P	19	34.80	27.4
R	24	40.29	31.7
PR	456	52.04	40.9
Total	499		100

As can be seen from Table 18, the residual component (PR) was the largest source of variance with 40.9%, which shows that a greater variance source cannot be explained because of the interplay between raters, persons, and other unexplained sources of error. Secondly, the rater facet (R) was the next biggest variance source with 31.7%, which signified a large proportion of inconsistent scores assigned by the raters for low-quality L2 speaking performances. Persons (P), whose performances seemed to show substantial differences, was the least source of variance with 27.4%. This could be related to the target of this analysis since a homogeneous design was used. However, the person-by-rater-by-quality (p *x* r *x* q) was a heterogeneous design including all speaking performance qualities, the proportion of persons (P) variance was comparatively larger than this person-by-rater (p *x* r) design of low-quality L2 speaking performances.

Similarly, the person-by-rater (p *x* r) random effects G-study design was conducted to reveal the sources of variation that contribute to the scores awarded to medium-quality L2 speaking performances. Table 19 illustrates the sources of variance that contribute to the score variability of medium-quality L2 speaking performances.

Table 19

Analysis of variance for random effects p x r design (medium-quality L2 speaking performances)

Source of Variance	*df*	σ^2	%
P	19	24.92	28.5
R	24	17.11	19.6
PR	456	45.35	51.9
Total	499		100

As shown in Table 19, the residual component (PR) was the largest variance with 51.9% due to the interaction between raters, persons, and other systematic and unsystematic sources of error. The second biggest source of variance was persons (P) with 28.5%, which means that the students whose speaking abilities were moderate showed differences. Finally, the least variance component was raters (R) with 19.6%, indicating that raters showed substantial differences while grading medium-quality speaking performances.

To examine the sources of variation that contribute to the scores awarded to high-quality L2 speaking performances, the person-by-rater (p *x* r) random effects G-study design was carried out. Table 20 provides information on the variance components that contribute to the score variability of high-quality L2 speaking performances.

Table 20

Analysis of variance for random effects p x r design (high-quality L2 speaking performances)

Source of Variance	*df*	σ^2	%
P	19	21.08	26.5
R	24	10.57	13.3
PR	456	47.80	60.2
Total	499		100

As presented in Table 20, similar to the trends observed in low-quality and medium-quality L2 speaking performances, the residual component (PR) was the largest with 60.2%, which refers to the interaction between persons, raters, and other sources of error that cannot be explained. Following this, persons (P) were the second largest variance component with 26.5%, which means that students showed differences in their performances. Finally, raters (R) were the smallest variance component with 13.3%, indicating that raters differed markedly while grading high quality speaking performances.

All in all, raters were relatively more consistent when all speaking performance qualities were taken into consideration. However, while assessing low-quality speaking performances, raters showed considerable differences more than they did in grading medium-quality, and high-quality speaking performances (31,7%, 19.6%, and 13.3%, respectively). This result could be related to students' differences in their low-quality, medium-quality, and high-quality speaking performances (27.4%, 28.5%, and 26.5%, respectively). Given that various interactions between raters, persons, and speaking performance quality were observed, it can be claimed that raters employed more different scoring patterns while grading low-quality speaking performances, yet they applied more similar scorings for medium-, and high-quality speaking performances.

Generalizability coefficient (Ep^2) and dependability coefficient (Φ) indices were formed to observe assessment situations across the person-by-rater-by-quality (p *x* r *x* q) design for all L2 speaking performances, and the person-by-rater (p *x* r) designs for low-, medium-, and high-quality L2 speaking performances. Table 21 summarizes the coefficients for all L2 speaking performances, and three speaking performance quality groups.

Table 21

Generalizability and dependability coefficients for speaking performance ratings

L2 speaking performances	$N_{Responses}$	N_{Raters}	Ep^2	Φ
All speaking performances	60	25	.98	.98
Low-quality speaking performances	20	25	.94	.90
Medium-quality speaking performances	20	25	.93	.91
High-quality speaking performances	20	25	.92	.90

Table 21 illustrates that the analysis provided higher generalizability and dependability coefficients for all L2 speaking performances (.98 and .98) than low-quality (.94 and .90), medium-quality (.93 and .91), and high-quality L2 speaking performances (.92 and .90). At the same time, there seems to be similarities of coefficient indices across low-, medium-, and high-quality L2 speaking performances.

In addition to the analyses of Generalizability coefficient (Ep^2) and Dependability coefficient (Φ), D-studies were computed both for the person-by-rater-by-quality *(p x r x q)* and the person-by-rater (p *x* r) crossed designs to create optimum assessment conditions. As higher generalizability coefficient (Ep^2) and dependability coefficient (Φ) figures refer to more convenient coefficient indices, the number of raters were formed accordingly. The dependability coefficients (Φ) of all crossed designs were higher than the acceptable level, which should be above .80. Therefore, the decreasing pattern in which the number of raters was reduced was utilized for this analysis. Table 22 presents the findings of D-study analysis that consists of generalizability and dependability coefficients for each crossed design.

Table 22

Generalizability and dependability coefficients for all, low-, medium-, and high quality speaking performances

All L2 speaking performances (*N*= 60, *p x r x q)*	N_{Raters}	Ep^2	Φ
	25	.98	.98
	20	.97	.97
	15	.97	.96
	10	.95	.94
	5	.90	.89
	3	**.85**	**.83**
	2	.79	.76
Low-quality L2 speaking performances (*N*= 20, *p x r)*	N_{Raters}	Ep^2	Φ
	25	.94	.90
	20	.93	.88
	15	.91	.85
	11	**.88**	**.81**
	10	.87	.79
	5	.77	.65
	3	.67	.53
Medium-quality L2 speaking performances (*N*= 20, *p x r)*	N_{Raters}	Ep^2	Φ
	25	.93	.91
	20	.92	.89
	15	.89	.86
	11	**.86**	**.81**
	9	.83	.78
	5	.73	.67
	3	.62	.54
High-quality L2 speaking performances (*N*= 20, *p x r)*	N_{Raters}	Ep^2	Φ
	25	.92	.90
	20	.90	.88
	15	.87	.84
	12	**.84**	**.81**
	10	.82	.78
	5	.69	.64
	3	.57	.52

Given that the minimum coefficient indices should be .80, in the first scenario of all L2 speaking performances (*p x r x q),* both Ep^2 and Φ coefficient figures were within the acceptable range. In fact, once the number of raters in all L2 speaking performances was reduced from 25 to 3, the Φ coefficient index decreased from .98 to .83. As for the scenario

for the low-quality L2 speaking performances, the last optimum number of raters was 11 and the Ep^2 and Φ coefficient indices were .88 and .81. Similarly, when the number of raters was decreased from 25 to 11, the Ep^2 and Φ coefficient indices would still give acceptable results with .86 and .81. Finally, if I decreased the number of raters for high-quality L2 speaking performances from 25 to 12, the coefficient indices would still be in the acceptable range (.84 and .81).

4.3.4. Results for RQ4

RQ4: Does the reliability (e.g., dependability coefficients for criterion-referenced score interpretations) of the analytic scores of raters (low, medium and high) differ from each other?

G-study tests were computed to analyze the level of variation across low-, medium-, and high-experienced raters whose scores for all speaking performance qualities, low-quality, medium-quality, and high-quality L2 speaking performances. Starting with the person-by-rater-by-quality *(p x r x q)* design, Table 23 summarizes the findings of coefficients for all qualities.

Table 23

Generalizability and dependability coefficients for all speaking performance qualities

Rater Experience Groups	N_{Raters}	$N_{SpResponses}$ (60)	Ep^2	Φ
Low-experienced	10		.96	.95
Medium-experienced	7	All Qualities	.95	.94
High-experienced	8		.94	.94

According to Table 23, all rater experience groups provided high Ep^2 and Φ coefficient indices, which refers to a high degree of concordance among rater experience

groups. As regards to the person-by-rater *(p x r)* design, various G-study tests were carried out to examine the coefficients for low-quality, medium-quality, and high-quality L2 speaking performances among separate rater groups. Table 24 provides information on the Ep^2 and Φ coefficient indices for low-quality L2 speaking performances.

Table 24

Generalizability and dependability coefficients for low-quality speaking performance scores

Rater Experience Groups	N_{Raters}	$N_{SpResponses}$ (20)	Ep^2	Φ
Low-experienced	10		.88	.80
Medium-experienced	7	Low-Quality	.83	.68
High-experienced	8		.77	.70

In Table 24, there were higher G-coefficients (Ep^2) for low-experienced and medium experienced raters (.88 and .83) than for the high-experienced ones (.77). Given the findings of dependability coefficients (Φ), the lowest figures were revealed for medium-, and high-experienced raters (.68 and .70). All in all, it seems that low-experienced raters produced higher Ep^2 and Φ coefficients (.88 and 80) than medium-, and high-experienced raters (.83 and .68; .77 and .70, respectively). Following that, Table 25 summarizes the Ep^2 and Φ coefficients for medium-quality L2 speaking performances.

Table 25

Generalizability and dependability coefficients for medium-quality speaking performance scores

Rater Experience Groups	N_{Raters}	$N_{SpResponses}$ (20)	Ep^2	Φ
Low-experienced	10		.83	.78
Medium-experienced	7	Medium-Quality	.86	.80
High-experienced	8		.78	.73

As can be seen in Table 25, there were higher G-coefficients for low-experienced and medium-experienced raters (.83 and .86) than high-experienced raters (.78). As for the dependability coefficients, medium-experienced raters provided higher coefficient (.80) than low-experienced and high-experienced raters (.78 and .73). Looking at general findings, I can say that medium-experienced raters had higher Ep^2 and Φ coefficients (.86 and .80) than the other two experience groups. Table 26 illustrates the summarizes the Ep^2 and Φ coefficients for high-quality L2 speaking performances.

Table 26

Generalizability and dependability coefficients for high-quality speaking performance scores

Rater Experience Groups	N_{Raters}	$N_{SpResponses}$ (20)	Ep^2	Φ
Low-experienced	10		.82	.78
Medium-experienced	7	High-Quality	.76	.69
High-experienced	8		.84	.84

Table 26 shows that higher Ep^2 coefficients were observed for high-experienced and low-experienced raters (.84 and .82) than medium-experienced raters (.76). As regards to showing the most consistent and inconsistent rater groups while scoring high-quality L2

speaking performances, it is clear that high-experienced raters had the highest G-, and dependability coefficients (.84 and .84). However, medium-experienced raters had the lowest coefficients (.76 and .69).

Utilizing the person-by-rater-by-quality *(p x r x q)* and person-by-rater *(p x r)* designs, G-coefficients and dependability coefficients for all speaking performance qualities, low-quality, medium-quality, and high-quality L2 speaking performances were examined in detail. Following that, D-studies were conducted to reveal the most ideal rating scenarios by manipulating the number of raters for each rater experience group. Table 27 shows the results of G-coefficient and dependability coefficients for low-experienced raters across four different scenarios as regards to decision studies framework.

Table 27

Generalizability and dependability coefficients for low-experienced raters

All L2 speaking performances (*N*= 60, *p x r x q)*	N_{Raters}	Ep^2	Φ
	10	.96	.95
	9	.94	.93
	5	.90	.88
	4	.88	.86
	3	**.85**	**.82**
	2	.79	.75
Low-quality L2 speaking performances (*N*= 20, *p x r)*	N_{Raters}	Ep^2	Φ
	10	.88	.80
	11	**.89**	**.81**
	12	.90	.83
	15	.92	.86
	18	.93	.88
	25	.95	.91
Medium-quality L2 sp performances (*N*= 20, *p x r)*	N_{Raters}	Ep^2	Φ
	10	.83	.78
	12	**.85**	**.81**
	15	.88	.84
	18	.90	.86
	25	.92	.90
	30	.94	.91
High-quality L2 sp performances (*N*= 20, *p x r)*	N_{Raters}	Ep^2	Φ
	10	.82	.78
	12	**.85**	**.81**
	15	.88	.84
	18	.89	.86
	25	.92	.90
	30	.93	.91

As can be seen in Table 27, even if the number of raters was decreased from 10 to three, the dependability coefficient for all L2 speaking performances would still be in the range of acceptable level (Φ = .82). As for low-quality L2 speaking performances if the number of raters was increased from 10 to 11, the dependability coefficient would reach an acceptable degree of index (Φ = .81). Similarly, when the number of raters was increased from 10 to 12 for medium-, and high-quality L2 speaking performances, the results would still give acceptable level of dependability coefficients (.81 and .81, respectively). Table 28

illustrates the Ep^2 and Φ coefficients for medium-experienced raters within the D-studies framework.

Table 28

Generalizability and dependability coefficients for medium-experienced raters

All L2 speaking performances (*N*= 60, *p x r x q)*	N_{Raters}	Ep^2	Φ
	7	.95	.94
	6	.93	.91
	5	.91	.90
	4	.90	.87
	3	**.87**	**.84**
	2	.81	.78
Low-quality L2 sp. performances (*N*= 20, *p x r)*	N_{Raters}	Ep^2	Φ
	7	.83	.68
	8	.85	.70
	10	.87	.75
	12	.89	.78
	14	**.91**	**.81**
	20	.93	.86
Medium-quality L2 sp. performances (*N*= 20, *p x r)*	N_{Raters}	Ep^2	Φ
	7	.86	.80
	8	**.88**	**.82**
	13	.92	.88
	16	.94	.90
	20	.95	.92
	25	.96	.94
High-quality L2 sp. performances (*N*= 20, *p x r)*	N_{Raters}	Ep^2	Φ
	7	.76	.69
	10	.82	.76
	14	**.86**	**.81**
	20	.90	.86
	25	.92	.89
	30	.93	.90

In Table 28, as for all L2 speaking performances, a total number of three medium-experienced raters would yield an acceptable degree of dependability coefficient index (Φ = .84). However, when the number of raters was manipulated from seven to 14 for low-quality L2 speaking performances, the dependability coefficient index was above the acceptable

degree of .80 (Φ = .81). In the same vein, the dependability coefficient would yield an acceptable index if the number of raters was increased from seven to 14 for high-quality L2 speaking performances (Φ = .81). Compared to low-quality and high-quality L2 speaking performances, the coefficient indices of medium-quality L2 speaking performances were more consistent. When the number of raters was increased just from seven to eight, the acceptable index would be reached (Φ = .82). Within the D-studies framework, Table 29 shows the Ep^2 and Φ coefficients for high-experienced raters across all L2 speaking performances, low-quality, medium-quality, and high-quality L2 speaking performances.

Table 29

Generalizability and dependability coefficients for high-experienced raters

All L2 speaking performances (*N*= 60, *p x r x q)*	N_{Raters}	Ep^2	Φ
	8	.94	.94
	7	.92	.92
	6	.91	.90
	4	.87	.86
	3	**.83**	**.82**
	2	.77	.76
Low-quality L2 speaking performances (*N*= 20, *p x r)*	N_{Raters}	Ep^2	Φ
	8	.77	.70
	12	.84	.78
	15	**.86**	**.81**
	20	.89	.85
	25	.91	.88
	30	.93	.90
Medium-quality L2 sp. performances (*N*= 20, *p x r)*	N_{Raters}	Ep^2	Φ
	8	.78	.73
	9	.80	.75
	12	.84	.80
	13	**.85**	**.81**
	20	.90	.87
	25	.92	.89
High-quality L2 speaking performances (*N*= 20, *p x r)*	N_{Raters}	Ep^2	Φ
	8	.84	.84
	7	**.82**	**.82**
	6	.80	.79
	5	.77	.76
	3	.67	.66
	2	.57	.56

As can be seen in Table 29, if the number of raters was decreased from eight to three for all L2 speaking performances, the dependability coefficient would still yield an acceptable level of index (Φ = .82). As regards to low-quality L2 speaking performances, a total number of 15 raters would be sufficient for an acceptable index (Φ = .81). A similar index was achieved for medium-quality L2 speaking performances with the number of 13 raters (Φ = .81). Finally, a total number of 7 raters would be needed for high-quality L2 speaking performances to achieve an acceptable level of dependability index (Φ = .82).

4.4. Qualitative Data Analysis Results

Verbal protocols and written score explanations were two principal components of the qualitative data in this study. A total number of 15 verbal protocols were completed by each rater. That is to say, the researcher allocated five verbal protocols to each speaking performance quality. The raters were not informed about this speaking performance quality division. Although sub-themes were determined inductively, the major categories and foci of the coding scheme used in this study were adapted from Cumming et al. (2002). The coding scheme included two main categories: a) interpretation strategies, and b) judgment strategies. Additionally, there were three main foci: a) self-monitoring-focus, b) rhetorical focus, and c) language-focus. Finally, the coding scheme was formed by 24 individual strategies under each main category and focus. Utilizing an inductive approach, the researcher obtained 15 major themes from the data set of written score explanations.

4.4.1. Findings for RQ5

RQ5: How do raters make decisions while rating varying quality L2 speaking performances analytically?

This research question aims to focus on examining how raters give decisions whilst scoring different quality L2 speaking performances. The tables provided in this part contains analyses presenting the strategies employed by raters in each speaking performance quality. To show the overall percentages of strategies used by the raters in three different quality L2 speaking performances, descriptive statistics were computed. Table 30 describes the main categories of decision-making behaviors reported by all rater groups across varying quality L2 speaking performances.

Table 30

Comparison of raters' decision-making behaviors across speaking performance quality

	Low-quality		**Medium-quality L2**		**High-quality L2**	
	Mdn	**Range**	***Mdn***	**Range**	***Mdn***	**Range**
Focus						
Self-Monitoring	19.81	6.67-36.00	20.83	10.53-34.67	22.94	4.65-36.67
Rhetorical	25.00	16.00-42.45	31.62	16.67-65.22	29.36	15.38-46.51
Language	51.28	39.02-66.67	48.57	23.91-63.04	48.39	28.57-63.15
Strategy						
Interpretation	6.76	0.00-51.45	6.67	0.00-54.17	7.34	0.00-41.76
Judgment	93.24	48.55-100.0	93.33	45.83-100.0	92.66	58.24-100.0
Strategy × Focus						
Interpretation						
Self-monitoring	2.50	0.00-19.18	1.96	0.00-10.71	2.20	0.00-15.29
Rhetorical	0.00	0.00-16.18	0.00	0.00-28.57	0.00	0.00-18.24
Language	2.17	0.00-24.24	2.17	0.00-25.88	0.00	0.00-15.09
Judgment						
Self-monitoring	16.04	3.33-34.00	17.86	4.35-29.03	20.63	2.33-36.67
Rhetorical	24.59	16.00-41.51	30.00	13.69-48.21	28.57	15.38-45.05
Language	47.62	25.43-63.33	42.86	21.74-55.00	46.15	27.47-58.95

Note. $N = 25$ raters.

As can be seen from Table 30, the percentages of decision-making behaviors showed slight differences across low-, medium-, and high-quality L2 speaking performances. Looking at the strategy focus, I can say that language strategy focus was the most commonly employed strategy across low-, medium-, and high-quality L2 speaking performances (*Mdn* = 51.28%, *Mdn* = 48.57%, and *Mdn* = 48.39% respectively). However, self-monitoring focus was the least utilized strategy (*Mdn* = 19.81%, *Mdn* = 20.83%, and *Mdn* = 22.94%). It is clear that raters employed more judgment strategies while assessing low-quality L2 speaking performances (*Mdn* = 93.24%), medium-quality L2 speaking performances (*Mdn* = 93.33%), and high-quality L2 speaking performances (*Mdn* = 92.66%) than interpretation strategies. In addition, the comparison of interpretation and judgment strategies revealed that rhetorical interpretation focus was the least frequently used strategy across all speaking performance groups while more self-monitoring interpretation focus was relatively utilized in low-quality and high-quality L2 speaking performances (*Mdn* = 2.50%, and *Mdn* = 2.20%) than the medium-quality ones (*Mdn* = 1.96%). On top of that, more language interpretation focus

strategy was reported by raters in low-, and medium-quality L2 speaking performances than the high-quality L2 speaking performances. Finally, examining judgement focused strategies, I can observe that raters reported more self-monitoring judgement strategies for high-quality L2 speaking performances (*Mdn* = 20.63%) than they did for low-quality and medium-quality L2 speaking performances (*Mdn* = 16.04% and *Mdn* = 17.86% respectively), whereas they uttered more rhetorical focused judgment strategies while assessing medium-quality L2 speaking performances (*Mdn* = 30.00%) than they did in low-, and high-quality L2 speaking performances. Moreover, raters reported more language focused judgement strategies for low-, and high-quality L2 speaking performances (*Mdn* = 47.62% and *Mdn* = 46.15%) than they did for medium-quality L2 speaking performances (*Mdn* = 42.86%).

To examine whether there were any significant differences of the aforementioned categories of strategies, inferential statistics were computed. Table 31 presents the Kruskall-Wallis test results of decision-making behaviors in low-quality, medium-quality, and high-quality L2 speaking performances.

Table 31

Kruskall-Wallis test results for major categories of decision-making behaviors across speaking performance quality

Major categories	***H*** **(2, *n* = 75)**	***p***	**Low-quality *(Mdn)***	**Medium-quality *(Mdn)***	**High-quality *(Mdn)***
Focus					
Self-monitoring	0.56	.75	19.81	20.83	22.94
Rhetorical	5.91	.05	25.00	31.62	29.36
Language	3.51	.15	51.28	48.57	48.39
Strategy					
Interpretation	0.26	.87	6.76	6.67	7.34
Judgment	0.26	.87	93.24	93.33	92.66
Strategy × Focus					
Interpretation					
Self-monitoring	1.84	.39	2.50	1.96	2.20
Rhetorical	0.85	.65	0.00	0.00	0.00
Language	0.01	.99	2.17	2.17	0.00
Judgment					
Self-monitoring	1.34	.51	16.04	17.86	20.63
Rhetorical	4.38	.11	24.59	30.00	28.57
Language	3.67	.15	47.62	42.86	46.15

According to the findings of Kruskall-Wallis test, there were not any significant differences in in the percentages of strategies. As can be seen in the median figures, the reason of this result could be related to the slight differences across speaking performance quality groups. For instance, although raters seemed to utter more language focused strategies in low-quality L2 speaking performances (*Mdn* = 51.28%) than they did in medium-, and high-quality L2 speaking performances (*Mdn* = 48.57 and *Mdn* = 48.39%), this difference was not statistically significant.

Conducting Kruskall-Wallis tests on the main categories of raters' decision-making behaviors across three speaking performance qualities, I carried out further Kruskall-Wallis tests on the individual categories of strategies to reveal possible significant differences. Table 32 provides information on the individual items of self-monitoring focus interpretation

and judgment strategies reported by raters in low-, medium-, and high-quality L2 speaking performances.

Table 32

Kruskall-Wallis test results for self-monitoring strategies across speaking performance quality

Individual categories	***H*** **(2, *n* = 75)**	***p***	**Low-quality** ***(Mdn)***	**Medium-quality** ***(Mdn)***	**High-quality** ***(Mdn)***
Self-monitoring focus-Interpretation strategies					
Interpret spoken response prompt or test items (SMI1)	0.50	.77	0.00	0.00	0.00
Consider personal situation of the test takers (SMI2)	9.29	.01	0.63	0.00	0.00
Refer to scoring rubric (SMI3)	2.34	.31	0.00	0.00	0.00
Self-monitoring focus-Judgment strategies					
Evaluate responses in comparison with other benchmarks or responses (SMJ1)	2.42	.29	0.00	0.00	0.00
State overall performance of the test takers (SMJ2)	7.50	.02	0.90	0.00	1.23
State or revisit scoring (SMJ3)	0.63	.72	13.46	17.65	15.91

According to Table 32, the Kruskall-Wallis test revealed a statistically significant difference in the strategy labelled as SMI2 across low-, medium-, and high-quality L2

speaking performances [(Gr1, $n_{\text{low-quality L2 speaking performances}} = 25$; Gr2, $n_{\text{medium-quality L2 speaking performances}} = 25$; Gr3, $n_{\text{high-quality L2 speaking performances}} = 25$), $H\ (2, n = 75) = 9.29, p = .01$]. Raters reported this strategy more frequently in the low-quality L2 speaking performances (*Mdn* = 0.63%) than they did in the medium-quality and the high-quality L2 speaking performances. Similarly, the test revealed a statistically significant difference in the strategy called SMJ2 across three speaking performance quality groups, [$H\ (2, n = 75) = 7.50, p = .02$]. The strategy SMJ2 recorded more in the high-quality L2 speaking performances (*Mdn* = 1.23%) than the other two speaking performance quality groups. No statistically significant differences were found between the groups for the percentages of other self-monitoring focus interpretation and judgment strategies that raters reported.

Mann-Whitney *U* tests were conducted to examine which of the self-monitoring focused interpretation and judgment strategies (SMI2 and SMJ2) were statistically significant from each other across low-, medium-, and high-quality L2 speaking performances. Table 33 presents the Mann-Whitney *U* test results for the self-monitoring strategies: a) 'consider personal situation of test takers' (SMI2) and b) 'state overall performance of test-takers' (SMJ2).

Table 33

Mann-Whitney *U* test results for self-monitoring strategies across speaking performance quality

Strategy	Quality Groups	*n*	*Mdn*	*U*	*z*	*p*	*R*
Interpretation Strategies							
Consider personal situation of the test takers (SMI2)	Low	25	0.63	268.5	-0.93	.35	.13
	Medium	25	0.00				
Consider personal situation of the test takers (SMI2)	Low	25	0.63	185.0	-2.98	.003	.42
	High	25	0.00				
Consider personal situation of the test takers (SMI2)	Medium	25	0.00	221.0	-2.30	.02	.32
	High	25	0.00				
Judgment Strategies							
State overall performance of the test takers (SMJ2)	Low	25	0.90	234.5	-1.66	.10	.14
	Medium	25	0.00				
State overall performance of the test takers (SMJ2)	Low	25	0.90	258.5	-1.08	.27	.15
	High	25	1.23				
State overall performance of the test takers (SMJ2)	Medium	25	0.00	180.5	-2.71	.01	.38
	High	25	1.23				

For SMI2, a Mann-Whitney *U* test disclosed statistically different results between the low-quality (*Mdn* = 0.00%, *n* = 25) and high-quality L2 speaking performances [(*Mdn* = 0.00%, *n* = 25), *U* = 185.0, *z* = -2.98, *p* = .003, *r* = .42]. Similar to this, the test revealed statistically different results for SMI2 between the medium-quality (*Mdn* = 0.00%, *n* = 25) and high-quality L2 speaking performances [(*Mdn* = 0.00%, *n* = 25), *U* = 221.0, *z* = -2.30, *p*

= .02, r = .32]. Finally, for SMJ2, the test revealed statistically different results between the medium-quality (*Mdn* = 0.00%, n = 25) and high-quality L2 speaking performances [(*Mdn* = 1.23%, n = 25), U = 180.5, z = -2.71, p = .01, r = .38]. Mann-Whitney U tests revealed no other significant results for the other paired groups.

Kruskall-Wallis tests were performed to determine which of the rhetorical and ideational focus strategies were statistically significant from each other. Table 34 summarizes the results for the individual items of rhetorical and ideational focus strategies reported by raters in low-, medium-, and high-quality L2 speaking performances.

Table 34

Kruskall-Wallis test results for rhetorical and ideational focus strategies across speaking performance quality

Individual categories of decision-making behaviors	***H* (2, *n* = 75)**	***p***	**Low-quality *(Mdn)***	**Medium-quality *(Mdn)***	**High-quality *(Mdn)***
Rhetorical and Ideational focus-Interpretation Strategies					
Interpret vague or equivocal expressions (RFI1)	0.54	.76	0.00	0.00	0.00
Restate test takers' ideas or propositions (RFI2)	2.71	.25	0.00	0.00	0.00
Rhetorical and Ideational focus-Judgment Strategies					
Evaluate topic development (RFJ1)	11.07	.004	6.31	10.00	11.54
Evaluate task completion, content and relevance (RFJ2)	3.11	.21	16.22	17.09	11.94
Evaluate originality and creativity (RFJ3)	1.28	.52	0.00	0.00	0.00
Recognize unnecessary or verbose expressions (RFJ4)	0.24	.88	0.00	0.00	0.00
Evaluate organization of the response (RFJ5)	5.06	.07	0.00	0.00	2.33
Evaluate register of the test takers (RFJ6)	0.90	.63	0.00	0.00	0.00

As presented in Table 34, raters reported significantly more the strategy called "Evaluate topic development" (RFJ1) for high-quality L2 speaking performances (*Mdn* = 11.54%) than they did for low-quality L2 speaking performances (*Mdn* = 6.31%) and

medium-quality L2 speaking performances [(*Mdn* = 10.00%), *H* (2, *n* = 75) = 11.07, *p* = .004]. Furthermore, raters uttered more the strategy labelled as RFJ2 in medium-quality L2 speaking performances (*Mdn* = 17.09%) than they did in high-quality L2 speaking performances (*Mdn* = 11.94%) and low-quality L2 speaking performances (*Mdn* = 16.22%), whereas this difference was not statistically significant. At the same time, no statistically significant differences were revealed in the other components of rhetorical and ideational focused interpretation and judgment strategies.

Following the Kruskall-Wallis tests, Mann-Whitney *U* tests were carried to find out which of the rhetorical and ideational focused interpretation and judgment strategy (RFJ1) was statistically significant from each other across low-, medium-, and high-quality L2 speaking performances. Table 35 summarizes the findings for RFJ1 in three speaking performance quality groups.

Table 35

Mann-Whitney *U* test results for rhetorical and ideational focus strategies across speaking performance quality

Strategy	**Quality Groups**	***n***	***Mdn***	***U***	***z***	***p***	***r***
Judgement Strategies							
Evaluate topic development	Low	25	6.31	182.5	-2.52	.01	.35
	Medium	25	10.00				
Evaluate topic development	Low	25	6.31	149.5	-3.16	.002	.44
	High	25	11.54				
Evaluate topic development	Medium	25	10.00	292.5	-0.38	.69	.05
	High	25	11.54				

As can be seen from Table 35, Mann-Whitney *U* tests revealed statistically different results between the low-quality (*Mdn* = 6.31%, *n* = 25) and medium-quality L2 speaking performances [(*Mdn* = 10.00%, *n* = 25), *U* = 182.5, *z* = -2.52, *p* = .01, *r* = .35]. Moreover, the test provided statistically different results between the low-quality (*Mdn* = 6.31%, *n* = 25) and high-quality L2 speaking performances [(*Mdn* = 11.54%, *n* = 25), *U* = 149.5, *z* = -3.16, *p* = .002, *r* = .44]. There was no statistically different result for the other paired group.

A Kruskall-Wallis test revealed statistically significant differences in two of the language focused judgment strategies across the three speaking performance quality groups. Table 36 provides information on the Kruskall-Wallis test findings for the individual categories of language focus interpretation and judgment strategies in low-, medium-, and high-quality L2 speaking performances.

Table 36

Kruskall-Wallis test results for language focus strategies across speaking performance quality

Individual categories	***H*** **(2, *n* = 75)**	***p***	**Low-quality *(Mdn)***	**Medium-quality *(Mdn)***	**High-quality *(Mdn)***
Language focus-Interpretation strategies					
Group errors into types (LFI1)	0.29	.86	1.64	1.85	2.27
Rephrase responses for interpretation (LFI2)	2.21	.33	0.00	0.00	0.00
Language focus-Judgment Strategies					
Evaluate intelligibility of the response (LFJ1)	11.30	.004	0.00	0.00	0.00
Consider errors in terms of quantity and frequency (LFJ2)	0.87	.64	2.50	3.92	4.44
Evaluate fluency (LFJ3)	6.47	.03	16.67	11.59	15.38
Evaluate vocabulary (LFJ4)	2.73	.25	9.76	10.74	11.63
Rate overall language use (LFJ5)	0.38	.82	0.00	0.00	0.00
Evaluate accent or pronunciation (LFJ6)	1.89	.38	2.89	2.56	1.76
Evaluate grammar and sentence structures (LFJ7)	4.58	.10	11.11	10.14	8.89
Evaluate L1 use of the test takers (LFJ8)	1.38	.50	0.00	0.00	0.00

According to the information presented in Table 36, Kruskall-Wallis tests revealed that there were statistically significant differences in the language focused judgment strategies labelled as LFJ1 (p = .004) and LFJ3 (p = .03). As for the strategy "Evaluate fluency," raters tended to rely on it more in low-quality L2 speaking performances (*Mdn* = 16.67%) than they did in high-quality L2 speaking performances (*Mdn* = 15.38%) and medium-quality L2 speaking performances (*Mdn* = 11.59%). Although there were no statistically significant differences were revealed in the other components, it would be useful to present the tendencies. For instance, raters seemed to consider more errors in terms of quantity and frequency (LFJ2) in high-quality (*Mdn* = 4.44%) and medium-quality L2 speaking performances (*Mdn* = 3.92%) than did in low-quality L2 speaking performances (*Mdn* = 2.50%). A similar trend was observed in the strategy "Evaluate vocabulary," which raters used less frequently in low-quality L2 speaking performances (*Mdn* = 9.76%) than did in high-quality (*Mdn* = 9.76%) and medium-quality L2 speaking performances (*Mdn* = 10.74%).

Following the Kruskall-Wallis tests, Mann-Whitney *U* tests were performed to examine the statistically significant pairs. Table 37 summarizes the results for language focus strategies labelled as LFJ1 and LFJ3 across low-quality, medium-quality, and high-quality L2 speaking performances.

Table 37

Mann-Whitney U test results for language focus strategies across speaking performance quality

Strategies	Quality Groups	*n*	*Mdn*	*U*	*z*	*p*	*r*
Judgment Strategies							
Evaluate intelligibility of the response	Low	25	0.00	204.5	-2.53	.01	.35
	Medium	25	0.00				
Evaluate intelligibility of the response	Low	25	0.00	193.0	-2.86	.004	.40
	High	25	0.00				
Evaluate intelligibility of the response	Medium	25	0.00	299.0	-0.43	.66	.06
	High	25	0.00				
Evaluate fluency	Low	25	16.67	194.5	-2.29	.02	.32
	Medium	25	11.59				
Evaluate fluency	Low	25	16.67	193.0	-2.86	.11	.40
	High	25	15.38				
Evaluate fluency	Medium	25	11.59	237.0	-1.46	.14	.20
	High	25	15.38				

As for the strategy "Evaluate intelligibility of the response," Mann-Whitney *U* tests revealed that there were statistically significant results for the two quality groups: a) low-, and medium- quality L2 speaking performances [(*Mdn* = 0.00%, and *Mdn* = 0.00%, *U* = 204.5, *z* = -2.53, *p* = .01, *r* = .35)], and b) low-, and high-quality L2 speaking performances

[(*Mdn* = 0.00%, and *Mdn* = 0.00%, *U* = 193.0, *z* = -2.86, *p* = .004, *r* = .40)]. However, there was no statistically significant result for the pair of medium-, and high-quality L2 speaking performances. Furthermore, the tests disclosed statistically different findings for the strategy of 'Evaluate fluency' in the pair of low-, and medium-quality L2 speaking performances [(*Mdn* = 16.67%, and *Mdn* = 11.59%, *U* = 194.5, *z* = -2.29, *p* = .02, *r* = .32)].

After presenting the main and subcategories of decision-making behaviors that raters reported during verbal protocols while assessing low-, medium-, and high-quality L2 speaking performances, the percentages of median figures were calculated to show the most frequently used strategies across all three speaking performance groups. Table 38, Table 39, and Table 40 illustrate the most commonly used decision-making behaviors by low-quality, medium-quality, and high-quality L2 speaking performances, respectively. The tables rank orders the top 10 decision-making behaviors.

Table 38

Medians for the most frequently used decision-making behaviors by low-quality L2 speaking performances

Decision-Making Behaviors	***Mdn* (%)**
Evaluate fluency	16.67
Evaluate task completion, content and relevance	16.22
State or revisit scoring	13.46
Evaluate grammar and sentence structures	11.11
Evaluate vocabulary	9.76
Evaluate topic development	6.31
Evaluate accent or pronunciation	2.89
Consider errors in terms of quantity and frequency	2.50
Group errors into types	1.64
State overall performance of the test takers	0.90

According to the information given in Table 38, it can be seen that raters mostly relied on the strategies "Evaluate fluency" (*Mdn* = 16.67%), "Evaluate task completion, content and relevance" (*Mdn* = 16.22%), and "State or revisit scoring" (*Mdn* = 13.46%). In addition, the item "Evaluate grammar and sentence structures" was the following most commonly used strategy (*Mdn* = 11.11%). Looking at the data in more detail, while 6 of the

strategies were language focused oriented, only 2 of them were self-monitoring strategies and the rest of them were rhetorical and ideational focused strategies. Therefore, I can observe that raters tended to report more language focused strategies than self-monitoring and rhetorical and ideational focused strategies for low-quality L2 speaking performances.

Table 39

Medians for the most frequently used decision-making behaviors by medium-quality L2 speaking performances

Decision-Making Behaviors	***Mdn* (%)**
State or revisit scoring	17.65
Evaluate task completion, content and relevance	17.09
Evaluate fluency	11.59
Evaluate vocabulary	10.74
Evaluate grammar and sentence structures	10.14
Evaluate topic development	10.00
Consider errors in terms of quantity and frequency	3.92
Evaluate accent or pronunciation	2.56
Group errors into types	1.85
Interpret speaking performance prompt or test items	0.00

As can be seen from Table 39, the two most commonly used strategies by raters while assessing medium-quality L2 speaking performances were "State or revisit scoring," and "Evaluate task completion, content and relevance" [(*Mdn* = 17.65%, and *Mdn* = 17.09%, respectively)]. These figures were followed by the strategies "Evaluate fluency" (*Mdn* = 11.59%) and "Evaluate vocabulary" (*Mdn* = 10.74%). Furthermore, similar tendencies were observed in the strategies "Evaluate grammar and sentence structures" (*Mdn* = 10.14%) and "Evaluate topic development" (*Mdn* = 10.00%). As the percentages were examined carefully, it can be noticed that language focused strategies were the most frequently used by raters in medium-quality L2 speaking performances. At the same time, this figure showed parallelism with the overall trend of low-quality L2 speaking performances. Despite this similarity, there were also some differences. While raters uttered "Evaluate fluency" as the most commonly used strategy for low-quality L2 speaking performances (*Mdn* = 16.67%), they reported this strategy as the third frequently used one for medium-quality L2 speaking performances (*Mdn* = 11.59%). Another difference between low-quality and medium-

quality L2 speaking performances was in the strategy "State or revisit scoring" [(*Mdn* = 13.46%, and *Mdn* = 17.65%, respectively)].

Table 40

Medians for the most frequently used decision-making behaviors by high-quality L2 speaking performances

Decision-Making Behaviors	***Mdn* (%)**
State or revisit scoring	15.91
Evaluate fluency	15.38
Evaluate task completion, content and relevance	11.94
Evaluate vocabulary	11.63
Evaluate topic development	11.54
Evaluate grammar and sentence structures	8.89
Consider errors in terms of quantity and frequency	4.44
Evaluate organization of the response	2.33
Group errors into types	2.27
Evaluate accent or pronunciation	1.76

As for the high-quality L2 speaking performances, the most frequently used strategy was "State or revisit scoring" (*Mdn* = 15.91%), followed by "Evaluate fluency" (*Mdn* = 15.38%) and "Evaluate task completion, content and relevance" (*Mdn* = 11.94%). Similar to low-quality, and medium-quality L2 speaking performances, raters mostly reported language focused strategies for high-quality L2 speaking performances. Considering the commonalities of strategy use by raters, "Evaluate fluency," "Evaluate task completion, content and relevance," and "State or revisit scoring" were among the top three most frequently reported strategies across low-quality, medium-quality, and high-quality L2 speaking performances.

To justify the findings retrieved from verbal protocols, I asked raters to provide a total number of three written-score explanations while assessing each speaking performance. Table 41 summarizes the findings of the Kruskall-Wallis tests for written score explanations reported by raters across low-, medium-, and high-quality L2 speaking performances.

Table 41

Kruskall-Wallis test results for written score explanations across speaking performance quality

Score Explanation	*H* (2, *n* = 75)	*p*	Low-quality *(Mdn)*	Medium-quality *(Mdn)*	High-quality *(Mdn)*
Fluency	0.79	.67	26.67	23.33	26.67
Vocabulary	1.68	.43	18.33	20.00	21.67
Grammar Use	0.61	.73	18.33	18.33	16.67
Task Completion	0.12	.93	10.00	10.00	11.67
Topic Development	1.36	.50	5.00	6.67	5.00
Relevance	1.39	.49	3.33	3.33	5.00
Pronunciation	7.08	.02	3.33	1.67	1.67
Organization	1.08	.58	1.67	1.67	1.67
Sentence Variety	4.26	.11	1.67	1.67	0.00
Content	0.62	.73	0.00	0.00	0.00
Intelligibility	2.22	.32	0.00	0.00	0.00
L1 Use	0.16	.92	0.00	0.00	0.00
Overall Language Use	1.15	.56	0.00	0.00	0.00
Overall Performance	2.00	.36	0.00	0.00	0.00
Redundancy	0.60	.74	0.00	0.00	0.00

As presented in Table 41, raters significantly reported more "Pronunciation" for low-quality L2 speaking performances than (*Mdn* = 3.33%) than they did for medium-quality (*Mdn* = 1.67%) and high-quality L2 speaking performances [(*Mdn* = 1.67%), *H* (2, *n* = 75) = 7.08, *p* = .02]. Although there were not any other statistically significant differences across three spoken-responses quality groups, it would be useful to compare and contrast overall tendencies. The strategies "Evaluate fluency" and "Evaluate task completion, content, and relevance" were among top three most frequently used decision-making behaviors across all speaking performance quality groups. Similar to this finding, "Fluency" was the most commonly reported written score explanation with figures across low-quality (*Mdn* = 26.67%), medium-quality (*Mdn* = 23.33%), and high-quality L2 speaking performances (*Mdn* = 26.67%). This was followed by "Vocabulary", "Grammar use", and "Task completion", which in essence showed similar frequencies with the decision-making behaviors that raters reported whilst thinking aloud. Written score explanations analysis illustrated that raters prioritized language focused judgmental score explanations rather than

rhetorical and self-monitoring reasoning. All in all, the findings from the written score explanations corroborate the results obtained from verbal protocols.

Following the aforementioned analysis, Mann Whitney *U* tests were conducted to compare the differences between speaking performance quality groups for the written score explanation "Pronunciation," which Kruskall-Wallis tests revealed statistically significant. Table 42 summarizes the findings for each group of the L2 speaking performances.

Table 42

Mann-Whitney *U* test results for the written score explanations across speaking performance quality

Score Explanation	**Quality Groups**	***n***	***Mdn***	***U***	***z***	***p***	***r***
Pronunciation	Low	25	3.33	248.0	-1.28	.20	.18
	Medium	25	1.67				
Pronunciation	Low	25	3.33	177.5	-2.69	.01	.38
	High	25	1.67				
Pronunciation	Medium	25	1.67	248.0	-1.30	.19	.18
	High	25	1.67				

As for the written score explanation item "Pronunciation," Mann-Whitney *U* tests revealed that there was a statistically significant result between low-, and high-quality L2 speaking performances [(*Mdn* = 3.33%, and *Mdn* = 1.67%, *U* = 177.5, *z* = -2.69, *p* = .01, *r* = .38]. However, there were no statistically significant results for the pairs of low-quality, and medium-quality L2 speaking performances; medium-quality, and high-quality L2 speaking performances.

In addition, thematic content analysis of written score explanations were carried out to determine positive and negative aspects that raters attributed while reasoning their scores.

Table 43 displays the breakdown of written score explanations into positive and negative meanings across low-quality, medium-quality, and high-quality L2 speaking performances. Except for the reasons “Content,” “Intelligibility,” “L1 Use,” “Overall Language Use,” “Overall Performance,” “Redundancy,” “Relevance positive,” “Sentence variety positive,” the rest of the written score explanations provided statistically significant differences.

Table 43

Kruskall-Wallis test results for the positive and negative written score explanations across speaking performance quality

Positive and Negative Score Explanations	***H*** **(2, *n* = 75)**	***p***	**Low-quality *(Mdn)***	**Medium-quality *(Mdn)***	**High-quality *(Mdn)***
Fluency Positive	45.28	.00	1.67	11.67	20.00
Fluency Negative	46.27	.00	21.67	11.67	3.33
Vocabulary Positive	44.30	.00	1.67	6.67	20.00
Vocabulary Negative	36.09	.00	18.33	11.67	3.33
Grammar Use Positive	29.52	.00	0.00	3.33	8.33
Grammar Use Negative	20.67	.00	16.67	15.00	8.33
Task Completion Positive	16.06	.00	1.67	6.67	10.00
Task Completion Negative	34.53	.00	8.33	3.33	0.00
Topic Development Positive	16.22	.00	1.67	3.33	5.00
Topic Development Negative	11.56	.003	3.33	0.00	0.00
Pronunciation Positive	7.24	.02	0.00	0.00	0.00
Pronunciation Negative	20.54	.00	3.33	1.67	0.00
Organization Positive	9.93	.01	0.00	0.00	1.67
Organization Negative	13.53	.00	1.67	0.00	0.00
Relevance Negative	8.64	.01	0.00	0.00	0.00
Sentence Variety Negative	14.11	.00	0.00	0.00	0.00

As can be seen in Table 43, raters naturally tended to prefer negative reasons for low-quality L2 speaking performances while they provided more positive reasons for high-quality L2 speaking performances. For instance, as for the median results of low-quality L2 speaking performances for "Fluency," 21.67% of the written score explanations was negative, whereas only 1.67% of the reasons was positive. However, the figures of "Fluency

positive" (*Mdn* = 20.00%) were opposite of "Fluency negative" (*Mdn* = 3.33%). On the other hand, there was a similarity between "Fluency positive" (*Mdn* = 11.67%) and "Fluency negative" (*Mdn* = 11.67%) in medium-quality L2 speaking performances. Looking at the general trend of the data, it can be easily observed that raters significantly used more negative score explanations for low-quality L2 speaking performances than medium-quality and high-quality L2 speaking performances, which at the same time meant that raters provided more positive reasons for the latter speaking performance groups.

Follow-up Mann-Whitney *U* tests were computed to compare the figures between the speaking performance quality groups for the positive and negative connotations of written score explanations. The statistically significant differences that the Mann-Whitney *U* tests revealed are summarized for each speaking performance pair in Table 44, Table 45 and Table 46.

Table 44

Mann-Whitney *U* test results for the positive and negative written score explanations between low and medium-quality L2 speaking performances

Positive and Negative Score Explanations	**Quality Groups**	***n***	***Mdn***	***U***	***z***	***p***	***r***
Fluency Positive	Low	25	1.67	65.5	-4.82	.00	.68
	Medium	25	11.67				
Fluency Negative	Low	25	21.67	80.0	-4.53	.00	.64
	Medium	25	11.67				
Grammar Use Positive	Low	25	0.00	190.0	-2.54	.01	.35
	Medium	25	3.33				
Organization Positive	Low	25	0.00	227.0	-1.97	.04	.27
	Medium	25	0.00				
Organization Negative	Low	25	1.67	206.0	-2.32	.02	.32
	Medium	25	0.00				
Task Completion Positive	Low	25	1.67	164.5	-2.91	.004	.41
	Medium	25	6.67				
Task Completion Negative	Low	25	8.33	116.0	-3.84	.00	.54
	Medium	25	3.33				
Topic Development Positive	Low	25	1.67	169.0	-2.87	.004	.40
	Medium	25	3.33				
Vocabulary Positive	Low	25	1.67	96.5	-3.03	.00	.42
	Medium	25	6.67				
Vocabulary Negative	Low	25	18.33	156.5	-4.23	.002	.59
	Medium	25	11.67				

According to the information provided in Table 44, although there was a statistically significant difference for "Grammar Use Positive" results between the low-quality (*Mdn* = 0.00%, *n* = 25) and medium-quality L2 speaking performances [(*Mdn* = 3.33%, *n* = 25), *U* = 190.0, *z* = -2.54, *p* = .01, *r* = .35], there was no statistically significant difference for "Grammar Use Negative" between low-, and medium-quality L2 speaking performances [(*Mdn* = 16.67%, and *Mdn* = 15.00%, respectively)]. Unlike "Topic Development Positive" [(*Mdn* = 1.67%, *n* = 25, and *Mdn* = 3.33%, *n* = 25), *U* = 169.0, *z* = -2.87, *p* = .004, *r* = .40], no statistically different result was observed for "Topic Development Negative" between low-quality (*Mdn* = 3.33%), and medium-quality L2 speaking performances (*Mdn* = 0.00%). The test revealed no statistically significant findings for the other positive and negative connotation pairs such as "Pronunciation Positive and Negative," "Relevance Negative," and "Sentence Variety Negative." In the following part, Table 45 displays Mann-Whitney *U* test results for positive and negative reasons between low-, and high-quality L2 speaking performances.

Table 45

Mann-Whitney *U* test results for the positive and negative written score explanations between low and high-quality L2 speaking performances

Written Score Explanations	**Quality Groups**	***n***	***Mdn***	***U***	***z***	***p***	***r***
Fluency Positive	Low	25	1.67	13.0	-5.83	.00	.82
	High	25	20.00				
Fluency Negative	Low	25	21.67	10.0	-5.89	.00	.83
	High	25	3.33				
Grammar Use Positive	Low	25	0.00	51.5	-5.17	.00	.73
	High	25	8.33				
Grammar Use Negative	Low	25	16.67	106.0	-4.02	.00	.56
	High	25	8.33				
Organization Positive	Low	25	0.00	168.0	-3.13	.002	.44
	High	25	1.67				
Organization Negative	Low	25	1.67	159.5	-3.48	.00	.49
	High	25	0.00				
Pronunciation Positive	Low	25	0.00	201.5	-2.66	.01	.37
	High	25	0.00				
Pronunciation Negative	Low	25	3.33	102.0	-4.46	.00	.63
	High	25	0.00				
Relevance Negative	Low	25	0.00	206.5	-2.75	.01	.38
	High	25	0.00				
Sentence Variety Negative	Low	25	0.00	143.5	-3.63	.00	.51
	High	25	0.00				

Task Completion Positive	Low	25	1.67	121.0	-3.75	.00	.53
	High	25	10.00				
Task Completion Negative	Low	25	8.33	38.0	-5.40	.00	.76
	High	25	0.00				
Topic Development Positive	Low	25	1.67	117.5	-3.89	.00	.55
	High	25	5.00				
Topic Development Negative	Low	25	3.33	161.0	-3.24	.00	.45
	High	25	0.00				
Vocabulary Positive	Low	25	1.67	2.5	-6.04	.00	.85
	High	25	20.00				
Vocabulary Negative	Low	25	18.33	29.5	-5.51	.00	.77
	High	25	3.33				

As can be seen from Table 45, Mann-Whitney *U* tests revealed statistically significant differences for all paired groups between low-quality and high-quality L2 speaking performances. Table 46 below illustrates the results for the positive and negative written score explanations between medium-, and high-quality L2 speaking performances.

Table 46

Mann-Whitney *U* test results for the positive and negative written score explanations between medium and high-quality L2 speaking performances

Positive and Negative Score Explanations	**Quality Groups**	***n***	***Mdn***	***U***	***z***	***p***	***r***
Fluency Positive	Medium	25	11.67	130.0	-3.54	.00	.50
	High	25	20.00				
Fluency Negative	Medium	25	11.67	106.5	-4.03	.00	.57
	High	25	3.33				
Grammar Use Positive	Medium	25	3.33	143.5	-3.30	.00	.46
	High	25	8.33				
Grammar Use Negative	Medium	25	15.00	118.5	-3.77	.00	.53
	High	25	8.33				
Relevance Negative	Medium	25	0.00	210.5	-2.66	.01	.37
	High	25	0.00				
Sentence Variety Negative	Medium	25	0.00	217.5	-2.19	.03	.30
	High	25	0.00				
Task Completion Negative	Medium	25	3.33	170.5	-2.87	.004	.40
	High	25	0.00				
Vocabulary Positive	Medium	25	6.67	128.5	-3.58	.00	.50
	High	25	20.00				
Vocabulary Negative	Medium	25	11.67	114.5	-3.87	.00	.54
	High	25	3.33				

As presented in Table 46, it can be seen that there was a statistically significant difference for "Task Completion Negative" results between the medium-quality (*Mdn* = 3.33%, *n* = 25) and high-quality L2 speaking performances [(*Mdn* = 0.00%, *n* = 25), *U* = 170.5, *z* = -2.87, *p* = .004, *r* = .40]. However, there was no statistically significant difference for "Task Completion Positive" between medium-, and high-quality L2 speaking performances [(*Mdn* = 6.67%, and *Mdn* = 10.00%, respectively)]. In addition, there were not any statistically different findings for the other paired reasons such as "Topic Development Positive," "Topic Development Negative," "Pronunciation Positive," "Pronunciation Negative," "Organization Positive," and "Organization Negative."

4.4.2. Findings for RQ6

RQ6: How does professional experience affect raters' decision-making processes and the aspects of speaking responses they focus on?

Table 47 compares the descriptive statistics of decision-making behaviors for each major category across low-experienced, medium-experienced, and high-experienced raters.

Table 47

Comparison of raters' decision-making behaviors across rater experience groups

	Low-experienced raters[a]		Medium-experienced raters[b]		High-experienced raters[c]	
	Mdn	**Range**	***Mdn***	**Range**	***Mdn***	**Range**
Focus						
Self-Monitoring	21.98	11.90-36.67	26.37	10.10-33.78	18.52	4.65-35.38
Rhetorical	29.58	16.00-43.48	26.32	16.67-45.05	29.74	15.38-65.22
Language	47.75	34.78-66.67	47.76	28.57-63.16	50.87	23.91-63.64
Strategy						
Interpretation	6.61	0.00-54.17	4.55	0.00-13.46	10.00	0.00-34.12
Judgment	93.39	45.83-100.0	95.45	86.54-100.0	90.00	65.88-100.0
Strategy × Focus						
Interpretation						
Self-monitoring	1.51	0.00-19.08	2.20	0.00-9.62	3.78	0.00-9.43
Rhetorical	0.00	0.00-28.57	0.00	0.00-3.75	0.78	0.00-25.88
Language	3.00	0.00-16.18	0.00	0.00-6.82	2.95	0.00-25.88
Judgment						
Self-monitoring	17.49	5.36-36.67	23.19	7.37-31.15	15.03	2.33-29.33
Rhetorical	26.71	13.69-42.86	26.25	16.67-45.05	29.19	15.38-48.21
Language	46.41	25.43-63.33	45.16	27.47-58.95	46.07	21.74-60.00

n^a = 10 raters. n^b = 7 raters. n^c = 8 raters

As Table 47 shows, "Language" focused strategies were the most frequently reported for the three rater groups (*Mdn* = 47.75%, 47.76%, and 50.87%). It is clear that high-experienced raters seemed to report slightly more language strategies than the low-, and medium experienced raters. The second commonly used decision-making behavior was "Rhetorical" focused strategies (*Mdn* = 29.58%, 26.32%, and 29.74%), which were quite similar to each other. These figures were followed by "Self-monitoring" focused strategies (*Mdn* = 21.98%, 26.37%, and 18.52%). The medium-experienced raters (*Mdn* = 26.37%) used more self-monitoring strategies than the other two experience groups. As for the category of interpretation and judgment, low-, medium-, and high-experienced rater groups reported more "Judgment" strategies (*Mdn* = 93.39%, 95.45%, and 90.0%) than "Interpretation" strategies (*Mdn* = 6.61%, 4.55%, and 10.00%).

To explore whether there were any significant differences among the major categories of strategies by rater experience groups, inferential statistics were conducted. Table 48 illustrates the Kruskall-Wallis test results of decision-making behaviors across low-, medium-, and high-experienced raters.

Table 48

Kruskall-Wallis test results for major categories of decision-making behaviors across rater experience groups

Major categories of decision-making behaviors	***H* (2, *n* = 25)**	***p***	**Low-experienced raters[a] *(Mdn)***	**Medium-experienced raters[b] *(Mdn)***	**High-experienced raters[c] *(Mdn)***
Focus					
Self-monitoring	8.79	.01	21.98	26.37	18.52
Rhetorical	3.36	.18	29.58	26.32	29.74
Language	1.35	.50	47.75	47.76	50.87
Strategy					
Interpretation	5.07	.07	6.61	4.55	10.00
Judgment	5.07	.07	93.39	95.45	90.00
Strategy × Focus					
Interpretation					
Self-monitoring	2.02	.36	1.51	2.20	3.78
Rhetorical	4.38	.11	0.00	0.00	0.78
Language	2.30	.31	3.00	0.00	2.95
Judgment					
Self-monitoring	9.19	.01	17.49	23.19	15.03
Rhetorical	1.85	.39	26.71	26.25	29.19
Language	1.06	.58	46.41	45.16	46.07

n^a = 10 raters. n^b = 7 raters. n^c = 8 raters

As can be seen from Table 48, there was a statistically significant finding in the percentage of "Self-monitoring" focused strategy [(Gr1, $n_{\text{low-experienced raters}}$ = 10; Gr2, $n_{\text{medium-experienced raters}}$ = 7; Gr3, $n_{\text{high-experienced raters}}$ = 8), H (2, n = 25) = 8.79, p = .01]. Similar to this, the test revealed a statistically significant difference in the "Self-monitoring" judgment strategy across three speaking performance quality groups, [H (2, n = 25) = 9.19, p = .01].

No statistically significant differences were found across the rater experience groups for the percentages of other major categories of decision-making behaviors.

Following the Kruskall-Wallis tests, Mann-Whitney *U* tests were carried to find out which of the self-monitoring focused and self-monitoring judgment strategies was statistically significant from each other across low-, medium-, and high-experienced raters. Table 49 summarizes the findings for these two strategy categories in three rater experience groups.

Table 49

Mann-Whitney *U* test results for major categories of decision-making behaviors across rater experience groups

Decision-making behaviors	**Experience Groups**	***n***	***Mdn***	***U***	***z***	***p***	***r***
Self-monitoring	Low	10	21.98	194.5	-2.30	.02	.55
	Medium	7	26.37				
Self-monitoring	Low	10	21.98	260.5	-1.73	.08	.40
	High	8	18.52				
Self-monitoring	Medium	7	26.37	147.0	-2.38	.02	.61
	High	8	18.52				
Self-monitoring Focus Judgment	Low	10	17.49	209.0	-2.02	.04	.49
	Medium	7	23.19				
Self-monitoring Focus Judgment	Low	10	17.49	267.0	-1.61	.10	.37
	High	8	15.03				
Self-monitoring Focus Judgment	Medium	7	23.19	431.0	-2.75	.01	.71
	High	8	15.03				

According to the information provided in Table 49, there was a statistically significant difference for "Self-monitoring" results between the low-experienced (*Mdn* = 21.98%, *n* = 10) and medium-experienced raters [(*Mdn* = 26.37%, *n* = 7), *U* = 194.5, *z* = -2.30, *p* = .02, *r* = .55]. Similarly, there was a statistically significant difference for "Self-monitoring" between medium-, and high-experienced raters [(*Mdn* = 26.37%, and *Mdn* = 18.52%, respectively), *U* = 147.0, *z* = -2.38, *p* = .02, *r* = .61]. However, there was no statistically significant difference for the pair of low-, and high-experienced raters. As for "Self-monitoring judgment" category, statistically significant differences can be observed for the pair of low-experienced raters (*Mdn* = 17.49%), and medium-experienced raters [(*Mdn* = 23.19%), *U* = 209.0, *z* = -2.02, *p* = .04, *r* = .49]. There was also a statistically significant difference between medium-experienced raters (*Mdn* = 23.19%), and high-experienced raters [(*Mdn* = 15.03%), *U* = 431.0, *z* = -2.75, *p* = .01, *r* = .71].

Following the findings of Kruskall-Wallis and Mann-Whitney *U* tests on the main categories of raters' decision-making behaviors, Kruskall-Wallis tests were conducted to examine the individual categories of strategies across three rater experience groups. Table 50 provides information on each self-monitoring focus interpretation and judgment strategy derived from verbal protocols across rater experience groups.

Table 50

Kruskall-Wallis test results for self-monitoring strategies across rater experience groups

Individual categories of decision-making behaviors	***H* (2, *n* = 25)**	***p***	**Low-experienced raters *(Mdn)***	**Medium-experienced raters *(Mdn)***	**High-experienced raters *(Mdn)***
Self-monitoring focus-Interpretation strategies					
Interpret spoken response prompt or test items (SMI1)	0.18	.91	0.00	0.00	0.00
Consider personal situation of the test takers (SMI2)	0.37	.82	0.00	0.00	0.00
Refer to scoring rubric (SMI3)	4.09	.12	0.00	0.00	1.12
Self-monitoring focus-Judgment strategies					
Evaluate responses in comparison with other benchmarks or responses (SMJ1)	5.33	.07	0.00	0.00	0.00
State overall performance of the test takers (SMJ2)	1.19	.55	0.85	0.00	0.91
State or revisit scoring (SMJ3)	8.91	.01	15.04	20.90	13.00

As shown in Table 50, the Kruskall-Wallis test revealed a statistically significant difference in the strategy labeled as SMJ3 across low-, medium-, and high-experienced raters, [H (2, n = 25) = 8.91, p = .01]. Medium-experienced raters reported this strategy more frequently (*Mdn* = 20.90%) than the low-experienced (*Mdn* = 15.04%) and high-experienced raters (*Mdn* = 13.00%). No statistically significant differences were found across the rater experience groups for the percentages of other self-monitoring strategies.

Mann-Whitney *U* tests were conducted to examine which pairs of the rater experience groups were statistically significant from each other as regards to "State or revisit scoring" (SMJ3). Table 51 summarizes the Mann-Whitney *U* test results for the aforementioned strategy.

Table 51

Mann-Whitney *U* Test Results for Self-Monitoring Strategies across Rater Experience Groups

Strategy	**Experience Groups**	***n***	***Mdn***	***U***	***z***	***p***	***r***
Judgment Strategies							
State or revisit scoring (SMJ3)	Low	10	15.04	193.5	-2.32	.02	.56
	Medium	7	20.90				
State or revisit scoring	Low	10	15.04	309.0	-0.88	.37	.20
	High	8	13.00				
State or revisit scoring	Medium	7	20.90	130.0	-2.77	.01	.71
	High	8	13.00				

As for the strategy "State or revisit scoring," a Mann-Whitney *U* test disclosed statistically different results between the low-experienced raters (*Mdn* = 15.04%, *n* = 10) and medium-experienced raters [(*Mdn* = 20.90%, *n* = 7), *U* = 193.5, *z* = -2.32, *p* = .02, *r* = .56]. Similarly, the test revealed statistically different results for this strategy between the medium-experienced (*Mdn* = 20.90%, *n* = 7) and high-experienced raters [(*Mdn* = 13.00%, *n* = 8), *U* = 130.0, *z* = -2.77, *p* = .01, *r* = .71].

Following that, Kruskall-Wallis tests were performed to determine which of the rhetorical and ideational focus strategies were statistically significant across rater experience

groups. Table 52 summarizes the results for the percentages of strategies by rhetorical and ideational focus items across low-, medium-, and high-experienced raters.

Table 52

Kruskall-Wallis test results for rhetorical and ideational focus strategies across rater experience groups

Individual categories	***H*** **(2, *n* = 25)**	***p***	**Low-experienced raters *(Mdn)***	**Medium-experienced raters *(Mdn)***	**High-experienced raters *(Mdn)***
Rhetorical and Ideational focus-Interpretation Strategies					
Interpret vague or equivocal expressions (RFI1)	7.42	.02	0.00	0.00	0.00
Restate test takers' ideas or propositions (RFI2)	1.48	.47	0.00	0.00	0.00
Rhetorical and Ideational focus-Judgment Strategies					
Evaluate topic development (RFJ1)	7.06	.03	6.51	8.70	12.86
Evaluate task completion, content and relevance (RFJ2)	2.65	.26	16.40	14.52	13.95
Evaluate originality and creativity (RFJ3)	0.38	.82	0.00	0.00	0.00
Recognize unnecessary or verbose expressions (RFJ4)	1.17	.55	0.00	0.00	0.00
Evaluate organization of the response (RFJ5)	0.05	.67	0.93	0.00	0.59
Evaluate register of the test takers (RFJ6)	1.84	.39	0.00	0.00	0.00

As presented in Table 52, the Kruskall-Wallis test revealed a statistically significant difference in the strategy labeled as RFI1 across low-, medium-, and high-experienced raters, [H (2, n = 25) = 7.42, p = .02]. Moreover, high-experienced raters reported more "Evaluate topic development" (RFJ1) strategy (Mdn = 12.86%) than the medium-experienced raters (Mdn = 8.70%) and low-experienced raters [(Mdn = 6.51%), H (2, n = 25) = 7.06, p = .03]. At the same time, no statistically significant differences were revealed in the other components of rhetorical and ideational focused interpretation and judgment strategies across three rater groups.

After performing Kruskall-Wallis tests, I carried out Mann-Whitney U tests to find out which rater experience pairs differentiated significantly from each other while reporting the strategies RFI1 and RFJ1. Table 53 summarizes the findings for RFI1 and RFJ1 across three rater experience groups.

Table 53

Mann-Whitney *U* test results for rhetorical and ideational focus strategies across rater experience groups

Strategies	Experience Groups	*n*	*Mdn*	*U*	*z*	*p*	*r*
Interpretation Strategies							
Interpret vague or equivocal expressions (RFI1)	Low	10	0.00	258.0	-1.72	.08	.41
	Medium	7	0.00				
Interpret vague or equivocal expressions	Low	10	0.00	295.0	-1.37	.17	.32
	High	8	0.00				
Interpret vague or equivocal expressions	Medium	7	0.00	164.0	-2.65	.01	.68
	High	8	0.00				
Judgment Strategies							
Evaluate topic development (RFJ1)	Low	10	6.51	277.0	-0.72	.46	.17
	Medium	7	8.70				
Evaluate topic development*	Low	10	6.51	216.5	-2.49	.01	.58
	High	8	12.86				
Evaluate topic development	Medium	7	8.70	168.0	-1.91	.05	.49
	High	8	12.86				

As can be seen from Table 53, Mann-Whitney *U* tests revealed a statistically different result for the interpretation strategy labelled as RFI1 between the medium-experienced (*Mdn* = 0.00%, *n* = 7) and high-experienced raters [(*Mdn* = 0.00%, *n* = 8), *U* = 164.0, *z* = -2.65, *p* = .01, *r* = .68]. Furthermore, as to the judgment strategy RFJ1, the test provided a statistically

different finding between the low-experienced raters (*Mdn* = 6.51%, *n* = 10) and high-experienced raters [(*Mdn* = 12.86%, *n* = 8), *U* = 216.5, *z* = -2.49, *p* = .01, *r* = .58]. There were no statistically significant results for the other paired groups.

A Kruskall-Wallis test revealed a statistically significant difference only in one of the language focused judgment strategies across three rater experience groups. Table 54 provides information on the Kruskall-Wallis test findings for the individual categories of language focus interpretation and judgment strategies in low-, medium-, and high-experienced raters.

Table 54

Kruskall-Wallis test results for language focus strategies across rater experience groups

Individual categories	***H*** **(2, *n* = 25)**	***p***	**Low-experienced raters** ***(Mdn)***	**Medium-experienced raters** ***(Mdn)***	**High-experienced raters** ***(Mdn)***
Language focus-Interpretation strategies					
Group errors into types (LFI1)	0.99	.60	2.20	1.77	2.00
Rephrase responses for interpretation (LFI2)	6.07	.048	0.00	0.00	0.00
Language focus-Judgment Strategies					
Evaluate intelligibility of the response (LFJ1)	1.56	.45	0.00	0.00	0.00
Consider errors in terms of quantity and frequency (LFJ2)	0.73	.69	4.00	3.85	2.24
Evaluate fluency (LFJ3)	3.33	.18	12.93	16.25	13.71
Evaluate vocabulary (LFJ4)	0.31	.85	10.56	10.62	10.13
Rate overall language use (LFJ5)	1.00	.60	0.00	0.00	0.00
Evaluate accent or pronunciation (LFJ6)	1.88	.39	2.27	1.10	2.94
Evaluate grammar and sentence structures (LFJ7)	1.67	.43	10.79	11.11	10.23
Evaluate L1 use of the test takers (LFJ8)	1.69	.43	0.00	0.00	0.00

According to the information presented in Table 54, the Kruskall-Wallis test revealed a statistically significant difference in the strategy labelled as LFI2 across low-, medium-, and high-experienced raters, [H (2, n = 25) = 6.07, p = .048]. There were not any statistically significant differences for the other language focus strategies across three rater groups.

Following the Kruskall-Wallis tests, Mann-Whitney U tests were conducted to examine the statistically significant pairs. Table 55 summarizes the results for the language focus strategy (LFI2) across low-, medium-, high-experienced raters.

Table 55

Mann-Whitney U test results for language focus strategies across rater experience groups

Strategies	**Experience Groups**	***n***	***Mdn***	***U***	***z***	***p***	***r***
Interpretation Strategies							
Rephrase responses for interpretation (LFI2)	Low	10	0.00	261.0	-1.49	.13	.36
	Medium	7	0.00				
Rephrase responses for interpretation	Low	10	0.00	300.0	-1.24	.21	.29
	High	8	0.00				
Rephrase responses for interpretation	Medium	7	0.00	169.0	-2.42	.02	.62
	High	8	0.00				

As shown in Table 55, Mann-Whitney U tests revealed a statistically significant result for the strategy labelled as LFI2 between the medium-experienced (Mdn = 0.00%, n = 7) and high-experienced raters [(Mdn = 0.00%, n = 8), U = 169.0, z = -2.42, p = .02, r = .62]. However, no statistically significant differences were revealed in the other pair of rater experience groups.

Table 56 compares the descriptive statistics of the percentages of decision-making behaviors derived from verbal protocols across rater experience groups and speaking performance quality. Furthermore, the data in this table provides the Kruskall-Wallis test results for statistically significant differences by speaking performance quality, and low-, medium-, and high-experienced raters.

Table 56

Comparison of main categories of decision-making behaviors by speaking performance quality and rater experience groups

Rater Group/Speaking Performance Quality	Low-quality			Medium-quality			High-quality		
	Mdn	**Min**	**Max**	***Mdn***	**Min**	**Max**	***Mdn***	**Min**	**Max**
Low-experienced[a]									
Focus									
Self-monitoring	21.69	13.33	36.00	18.93	14.77	27.45	22.58	11.90	36.67
Rhetorical	25.53	16.00	42.45	30.71	19.57	42.86	32.79	23.08	43.48
Language	51.61	39.02	66.67	47.62	39.22	63.04	44.44	34.78	58.93
Strategy									
Interpretation	4.26	0.00	51.45	6.90	1.37	54.17	6.90	0.00	41.76
Judgment	95.73	48.55	100.0	93.09	45.83	98.63	93.10	58.24	100.0
Strategy × Focus									
Interpretation									
Self-monitoring	2.02	0.00	19.08	1.69	0.00	10.71	0.00	0.00	15.29
Rhetorical	0.00	0.00	16.18	0.00	0.00	28.57	0.81	0.00	18.24
Language	2.02	0.00	16.18	3.78	0.00	14.88	4.15	0.00	8.89
Judgment									
Self-monitoring	13.81	6.36	34.00	17.72	5.36	25.49	20.87	7.65	36.67
Rhetorical	23.35	16.00	41.51	28.42	13.69	42.86	30.41	15.38	42.03
Language	49.23	25.43	63.33	44.76	26.79	55.00	41.25	28.99	52.98
Medium-experienced[b]									
Focus									
Self-monitoring	27.38	10.10	33.78	25.76	10.53	30.43	27.47	10.53	29.85
Rhetorical	25.00	19.57	32.69	26.39	16.67	38.05	26.32	17.58	45.05
Language	45.95	42.50	60.87	52.78	37.68	57.14	47.76	28.57	63.16
Strategy									
Interpretation	6.76	1.64	13.46	3.23	0.00	7.25	7.37	0.00	9.09
Judgment	93.24	86.54	98.36	96.77	92.75	100.0	92.63	90.91	100.0
Strategy × Focus									
Interpretation									
Self-monitoring	2.50	0.00	9.62	0.00	0.00	7.25	2.27	0.00	7.69
Rhetorical	0.00	0.00	3.75	0.00	0.00	1.77	0.00	0.00	1.49
Language	2.02	0.00	3.85	1.77	0.00	3.23	1.10	0.00	6.82
Judgment									
Self-monitoring	17.86	10.10	31.15	23.19	10.53	29.03	25.27	7.37	28.57

Rhetorical	25.00	18.92	32.69	26.39	16.67	36.28	26.32	17.58	45.05
Language	43.24	40.38	58.70	50.00	37.17	54.76	43.28	27.47	58.95
High-experienced[c]									
Focus									
Self-monitoring	18.70	6.67	34.67	16.62	10.71	34.67	17.50	4.65	35.38
Rhetorical*	24.66	18.18	33.85	33.86	25.33	65.22	30.27	15.38	46.51
Language	54.92	44.00	63.64	44.89	23.91	54.32	50.72	46.91	56.57
Strategy									
Interpretation	10.00	0.00	27.27	10.46	0.00	34.12	7.51	0.00	19.81
Judgment	90.00	72.73	100.0	89.53	65.88	100.0	92.48	80.19	100.0
Strategy × Focus									
Interpretation									
Self-monitoring	4.33	0.00	9.43	5.55	0.00	8.00	2.58	0.00	6.15
Rhetorical	0.45	0.00	3.33	1.11	0.00	23.91	0.94	0.00	6.98
Language	3.66	0.00	24.24	2.42	0.00	25.88	2.53	0.00	15.09
Judgment									
Self-monitoring	15.77	3.33	29.33	13.00	4.35	26.67	15.36	2.33	29.23
Rhetorical	24.21	17.17	33.85	32.51	21.18	48.21	29.19	15.38	39.53
Language*	47.40	39.39	60.00	37.18	21.74	50.68	48.15	36.79	50.98

n^a = 10 raters. n^b = 7 raters. n^c = 8 raters

* Kruskall-Wallis tests revealed that the differences across speaking perfornance qualities by rater experience groups were statistically significant at $p < .05$.

Looking at the general tendencies in Table 56, "Judgment" was the most frequently reported strategy by all rater groups across all three speaking performance quality types. Similarly, all rater groups reported more "Language" focused strategies than "Self-monitoring and Rhetorical" strategies for all speaking performance qualities. Low-experienced raters seemed to use more "Rhetorical-judgment" strategies for high-quality L2 speaking performances (*Mdn* = 30.41%) than medium-quality (*Mdn* = 28.42%) and low-quality L2 speaking performances (*Mdn* = 23.35%). At the same time, they reported more "Language-judgment" strategies for low-quality (*Mdn* = 49.23%) than medium-quality (*Mdn* = 44.76%) and high-quality L2 speaking performances (*Mdn* = 41.25%). Unlike low-experienced raters, medium-experienced raters showed similar figures for "Rhetorical-judgment" strategies across three speaking performance qualities. Furthermore, medium-experienced raters reported more "Language-judgment" strategies for medium-quality L2 speaking performances (*Mdn* = 50%) than low-quality (*Mdn* = 43.23%) and high-quality L2

speaking performances (*Mdn* = 43.28%). However, these differences were not statistically significant. High-experienced raters reported significantly more "Rhetorical" focus strategies for medium-quality L2 speaking performances (*Mdn* = 33.86%) than high-quality (*Mdn* = 30.27%) and low-quality L2 speaking performances [(*Mdn* = 24.66%), *H* (2, *n* = 8) = 6.04, *p* = .049]. Furthermore, high-experienced raters produced significantly less "Language-judgment" strategies for medium-quality L2 speaking performances (*Mdn* = 37.18%) than low-quality (*Mdn* = 47.40%) and high-quality L2 speaking performances [(*Mdn* = 48.15%), *H* (2, *n* = 8) = 7.38, *p* = .03]. There were not any other statistically significant differences across speaking performance quality by rater experience groups.

Following the Kruskall-Wallis tests, Mann-Whitney *U* tests were conducted to determine the statistically significant pairs. Table 57 summarizes the results for "Rhetorical" focus and "Language-judgment" strategies across speaking performance qualities within high-experienced raters.

Table 57

Mann-Whitney *U* test results for major categories of decision-making behaviors by speaking performance quality and high-experienced raters

Decision-making behaviors	Quality Groups	*n*	*Mdn*	*U*	*z*	*p*	*r*
Focus							
Rhetorical	Low	8	24.66	8.0	-2.52	.01	.63
	Medium	8	33.86				
Rhetorical	Low	8	24.66	19.0	-1.36	.19	.34
	High	8	30.27				
Rhetorical	Medium	8	33.86	24.0	-0.84	.44	.21
	High	8	30.27				
Strategy × Focus Judgment							
Language	Low	8	47.40	8.0	-2.52	.01	.63
	Medium	8	37.18				
Language	Low	8	47.40	29.0	-0.31	.79	.07
	High	8	48.15				
Language	Medium	8	37.18	12.0	-2.10	.04	.52
	High	8	48.15				

As presented in Table 57, it can be seen that there was a statistically significant difference for "Rhetorical" focus results between the low-quality (*Mdn* = 24.66%) and medium-quality L2 speaking performances [(*Mdn* = 33.86%), *U* = 8.0, *z* = -2.52, *p* = .01, *r* = .63]. Furthermore, as to "Language-judgment" strategy, the test provided statistically different findings between the low-quality L2 speaking performances (*Mdn* = 47.40%), and

medium-quality L2 speaking performances [(*Mdn* = 37.18%), *U* = 8.0, *z* = -2.52, *p* = .01, *r* = .63]; medium-quality L2 speaking performances (*Mdn* = 37.18%), and high-quality L2 speaking performances [(*Mdn* = 48.15%), *U* = 12.0, *z* = -2.10, *p* = .04, *r* = .52].

Table 58, Table 59, and Table 60 compare the top ten most frequently reported individual strategies by low-experienced, medium-experienced, and high-experienced raters while rating low-, medium-, and high-quality L2 speaking performances.

Table 58

The most common individual decision-making behaviors by speaking performance quality and low-experienced raters

Rater Group/Speaking Performance Quality	**Low-quality *(Mdn)***	**Medium-quality *(Mdn)***	**High-quality *(Mdn)***
Low-experienced raters			
Evaluate task completion, content and relevance (RFJ2)	15.88	17.54	15.84
State or revisit scoring (SMJ3)	12.68	17.26	14.64
Evaluate fluency (LFJ3)	17.19	10.80	13.07
Evaluate vocabulary (LFJ4)	10.07	11.25	11.41
Evaluate grammar and sentence structures (LFJ7)	11.55	9.98	8.79
Evaluate topic development (RFJ1)	5.37	7.78	10.15
Consider errors in terms of quantity and frequency (LFJ2)	2.85	5.22	4.72
Evaluate accent or pronunciation (LFJ6)	2.66	2.27	1.71
Group errors into types (LFI1)	0.61	3.78	2.86
Evaluate organization of the response (RFJ5)	0.46	1.01	2.20

As can be seen from Table 58, the most commonly used strategy for all L2 speaking performances by low-experienced raters was a rhetorical focused judgment strategy called

“Evaluate task completion, content and relevance.” This was followed by “State or revisit scoring,” “Evaluate fluency,” “Evaluate vocabulary,” and “Evaluate grammar and sentence structures.” Therefore, I can express that low-experienced raters generally tended to report language focused judgment strategies while rating low-, medium-, and high-quality L2 speaking performances. There were not any statistically significant differences across all speaking performance qualities by low-experienced raters.

Table 59

The most common individual decision-making behaviors by speaking performance quality and medium-experienced raters

Rater Group/Speaking Performance Quality	**Low-quality *(Mdn)***	**Medium-quality *(Mdn)***	**High-quality *(Mdn)***
Medium-experienced raters			
State or revisit scoring (SMJ3)	16.67	23.19	20.90
Evaluate fluency (LFJ3)	16.25	16.67	15.91
Evaluate task completion, content and relevance (RFJ2)	16.22	14.52	9.47
Evaluate vocabulary (LFJ4)	8.20	12.12	11.54
Evaluate grammar and sentence structures (LFJ7)	11.11	11.90	8.96
Evaluate topic development (RFJ1) *	6.73	8.70	11.36
Consider errors in terms of quantity and frequency (LFJ2)	4.05	1.52	3.85
Group errors into types (LFI1)	2.02	1.77	1.10
Evaluate accent or pronunciation (LFJ6)	0.00	3.03	1.10
Evaluate organization of the response (RFJ5)	0.00	0.00	2.99

In Table 59, the most common decision-making strategy was the self-monitoring judgment strategy "State or revisit scoring." "Evaluate fluency," and "Evaluate task completion, content and relevance" were the next two most frequently reported strategies. Despite some slight differences, I can observe certain similarities between low-experienced

and medium-experienced raters as regards to the rank of top frequently used strategies. Furthermore, the Kruskall-Wallis tests revealed statistically significant results for only one strategy. Medium-experienced raters reported more "Evaluate topic development" (RFJ1) strategy for high-quality L2 speaking performances (*Mdn* = 11.36%) than the medium-quality L2 speaking performances (*Mdn* = 8.70%) and low-quality L2 speaking performances [(*Mdn* = 6.73%), *H* (2, *n* = 21) = 6.58, *p* = .04]. Following Kruskall-Wallis tests, I conducted Mann-Whitney *U* tests to examine which pairs differentiated significantly from each other. According to the findings, there was a statistically significant result between the low-quality (*Mdn* = 6.73%, *n* = 7) and high-quality L2 speaking performances [(*Mdn* = 11.36%, *n* = 7), *U* = 6.0, *z* = -2.36, *p* = .02, *r* = .63]. There were no statistically significant differences for the other pairs.

Table 60

The most common individual decision-making behaviors by speaking performance quality and high-experienced raters

Rater Group/Speaking Performance Quality	**Low-quality *(Mdn)***	**Medium-quality *(Mdn)***	**High-quality *(Mdn)***
High-experienced raters			
Evaluate fluency (LFJ3)	15.51	10.47	16.33
Evaluate task completion, content and relevance (RFJ2)	16.02	14.95	9.73
State or revisit scoring (SMJ3)	14.23	13.00	12.79
Evaluate topic development (RFJ1)	6.49	15.48	13.17
Evaluate vocabulary (LFJ4)	9.95	8.48	12.26
Evaluate grammar and sentence structures (LFJ7)	11.50	9.26	8.43
Evaluate accent or pronunciation (LFJ6)	3.55	2.99	1.89
Consider errors in terms of quantity and frequency (LFJ2)	1.95	3.03	3.53
Group errors into types (LFI1)	2.66	1.77	2.53
Refer to scoring rubric (SMI3)	3.66	0.61	0.50

According to information presented in Table 60, it can be seen that “Evaluate fluency,” “Evaluate task completion, content and relevance,” “State or revisit scoring,” Evaluate topic development,” “Evaluate vocabulary” were the top five most frequently reported decision-making behaviors by high-experienced raters while rating low-, medium-

, and high-quality L2 speaking performances. Despite having certain similarities with low-, and medium-experienced raters, high-experienced raters preferred “Evaluate fluency” more than the other strategies. However, none of these differences were statistically significant.

Table 61 summarizes the findings of the Kruskall-Wallis tests for written score explanations reported by raters across low-, medium-, and high-quality L2 speaking performances.

Table 61

Kruskall-Wallis test results for the written score explanations across rater experience groups

Score Explanation	***H*** **(2, *n* = 25)**	***p***	**Low-experienced raters[a]** ***(Mdn)***	**Medium-experienced raters[b]** ***(Mdn)***	**High-experienced raters[c]** ***(Mdn)***
Fluency	1.31	.51	25.83	23.33	27.50
Vocabulary	10.7	.01	17.50	23.33	21.67
Grammar Use	12.0	.002	24.16	16.67	15.83
Task Completion	4.55	.10	6.67	15.00	10.00
Topic Development	5.41	.06	5.00	1.67	6.67
Relevance	7.86	.02	2.50	3.33	5.83
Pronunciation	0.94	.62	1.67	3.33	3.33
Organization	1.58	.45	1.67	3.33	1.67
Sentence Variety	1.08	.58	1.67	1.67	1.67
Intelligibility	7.37	.03	0.00	0.00	0.00
L1 Use	6.22	.045	0.00	0.00	0.00
Content	4.06	.13	0.00	0.00	0.00
Overall Language Use	2.18	.33	0.00	0.00	0.00
Overall Performance	1.50	.47	0.00	0.00	0.00
Redundancy	0.33	.84	0.00	0.00	0.00

n^a = 10 raters. n^b = 7 raters. n^c = 8 raters

As presented in Table 61, medium-experienced raters significantly reported more "Vocabulary" (*Mdn* = 23.33%) than high-experienced raters (*Mdn* = 21.67%), and low-experienced raters [(*Mdn* = 17.50%), *H* (2, *n* = 25) = 10.7, *p* = .01]. Another significant difference was observed in the item "Grammar Use." Low-experienced raters tended to attribute more "Grammar Use" to their score reasoning (*Mdn* = 24.16%) than medium-

experienced (*Mdn* = 16.67%), and high-experienced raters [(*Mdn* = 15.83%), *H* (2, *n* = 25) = 12.0, *p* = .002]. Furthermore, high-experienced raters significantly reported more "Relevance" (*Mdn* = 5.83%) than medium-experienced raters (*Mdn* = 3.33%), and low-experienced raters [(*Mdn* = 2.50%), *H* (2, *n* = 25) = 7.86, *p* = .02]. Finally, there were statistically significant differences for the score explanations "Intelligibility" and "L1 Use" across rater experience groups. Although there were not any statistically significant differences, it would be important to note that high-experienced raters reported more "Fluency" (*Mdn* = 27.50%) than low-experienced (*Mdn* = 25.83%), and medium-experienced raters (*Mdn* = 23.33%).

The aforementioned analysis showed that the written score explanations that all rater groups reported corroborate the results obtained from verbal protocols. For instance, the rater groups generally focused on strategies such as evaluating fluency, task completion, topic development, vocabulary, and grammar structures while forming the verbal protocols. Similar overall tendencies can be observed in the score explanations across three rater groups.

Following this, Mann Whitney *U* tests were conducted to compare the differences across rater experience groups for the written score explanations that Kruskall-Wallis tests revealed statistically significant. Table 62 summarizes the statistically significant findings for the relevant rater groups.

Table 62

Mann-Whitney *U* test results for the written score explanations across rater experience groups

Score Explanation	Rater Groups	*n*	*Mdn*	*U*	*z*	*p*	*r*
Grammar Use	Low	10	24.16	152.0	-3.63	.00	.85
	High	8	15.83				
Vocabulary	Low	10	17.50	164.0	-2.90	.004	.70
	Medium	7	23.33				
Vocabulary	Low	10	17.50	218.5	-2.74	.01	.64
	High	8	21.67				
Intelligibility	Medium	7	0.00	190.5	-2.22	.03	.57
	High	8	0.00				
Intelligibility	Low	10	0.00	252.5	-2.66	.008	.62
	High	8	0.00				
L1 Use	Medium	7	0.00	204.0	-2.21	.03	.57
	High	8	0.00				
L1 Use	Low	10	0.00	276.0	-2.50	.01	.58
	High	8	0.00				
Relevance	Medium	7	2.50	127.5	-2.86	.004	.73
	High	8	3.33				

According to the findings retrieved from Mann Whitney *U* tests, it was revealed that there was a statistically significant result between low-, and high-experienced raters for "Grammar Use" [(*Mdn* = 24.16%, and *Mdn* = 15.83%), *U* = 152.0, *z* = -3.63, *p* < .001, *r* = .85]. As for "Vocabulary," medium-, and high-experienced raters statistically reported more this explanation than low-experienced raters [(*Mdn* = 23.33%, and *Mdn* = 17.50%), *U* = 164.0, *z* = -2.90, *p* = .004, *r* = .70], and [(*Mdn* = 21.67%, and *Mdn* = 17.50%), *U* = 218.5, *z* = -2.74, *p* = .01, *r* = .64]. The tests also revealed statistically significant differences for "Intelligibility" and "L1 Use" between medium-, and high-experienced; low-, and high-experienced raters. Finally, high-experienced raters reported more "Relevance" (*Mdn* = 3.33%) than medium-experienced raters [(*Mdn* = 2.50%), *U* = 127.5, *z* = -2.86, *p* = .004, *r* = .73].

Table 63, Table 64, and Table 65 illustrate the descriptive statistics results of written score explanations that were classified as positive and negative meanings across low-experienced, medium-experienced, and high-experienced raters.

Table 63

Medians for the positive and negative written score explanations by low-experienced raters

Score Explanation	**Positive Reasons**	**Negative Reasons**
	***Mdn* (%)**	***Mdn* (%)**
Grammar Use	3.33	19.16
Fluency	8.33	13.33
Vocabulary	5.00	8.33
Task Completion	4.16	3.33
Topic Development	1.67	0.83
Pronunciation	0.00	1.67
Relevance	0.00	0.00
Organization	0.00	0.00
Sentence Variety	0.00	0.00
Content	0.00	0.00
Intelligibility	0.00	0.00
L1 Use	0.00	0.00
Overall Language Use	0.00	0.00
Overall Performance	0.00	0.00
Redundancy	0.00	0.00

As can be seen in Table 63, low-experienced raters generally tended to report more negative reasons for "Grammar Use," "Fluency," and "Vocabulary" (*Mdn*= 19.16%, *Mdn*= 13.33%, *Mdn*= 8.33%) than positive reasons (*Mdn*= 3.33%, *Mdn*= 8.33%, *Mdn*= 5.00%). However, as for "Task Completion" and "Topic Development," low-experienced raters used slightly more positive reasons (*Mdn*= 4.16%, *Mdn*= 1.67%) than negative reasons (*Mdn*= 3.33%, *Mdn*= 0.83%).

Table 64

Medians for the positive and negative written score explanations by medium-experienced raters

Score Explanation	**Positive Reasons**	**Negative Reasons**
	***Mdn* (%)**	***Mdn* (%)**
Fluency	11.67	13.33
Grammar Use	3.33	13.33
Vocabulary	11.67	11.67
Task Completion	6.67	5.00
Topic Development	1.67	0.00
Pronunciation	0.00	0.00
Relevance	1.67	0.00
Organization	3.33	0.00
Sentence Variety	0.00	1.67
Content	0.00	0.00
Intelligibility	0.00	0.00
L1 Use	0.00	0.00
Overall Language Use	0.00	0.00
Overall Performance	0.00	0.00
Redundancy	0.00	0.00

Compared to the figures by low-experienced raters, medium-experienced raters generally showed slight differences between the percentages of positive and negative reasons that they provided while rating L2 speaking performances. In fact, there was not a widening gap between two reasons in "Fluency," "Vocabulary," and "Task Completion." However, medium-experienced raters tended to report more negative reasons (*Mdn* = 13.33) than positive reasons (*Mdn* = 3.33) for "Grammar Use," which was similar to the figures by low-experienced raters.

Table 65

Medians for the positive and negative written score explanations by high-experienced raters

Written Score Explanation	**Positive Reasons**	**Negative Reasons**
	***Mdn* (%)**	***Mdn* (%)**
Fluency	10.83	10.83
Grammar Use	2.50	10.83
Vocabulary	10.83	8.33
Task Completion	6.67	2.50
Topic Development	3.33	1.67
Pronunciation	0.00	0.83
Relevance	5.00	0.00
Organization	0.00	0.00
Sentence Variety	0.00	0.00
Content	0.00	0.00
Intelligibility	0.00	0.00
L1 Use	0.00	0.00
Overall Language Use	0.00	0.00
Overall Performance	0.00	0.00
Redundancy	0.00	0.00

Looking at the overall findings, I can observe a similarity between medium-experienced and high-experienced rater groups, both of which tended to report positive and negative reasons with slight differences in comparison to low-experienced raters. Unlike other score explanations, high-experienced raters reported more negative reasons (*Mdn* = 10.83%) than positive reasons (*Mdn* = 2.50%) only for "Grammar Use." Table 65 below summarizes the statistically significant findings of the Kruskall-Wallis tests for written score explanations across low-, medium-, and high-experienced raters.

Table 66

Kruskall-Wallis test results for the positive and negative written score explanations across rater experience groups

Positive and Negative Score Explanations	***H*** **(2, *n* = 25)**	***p***	**Low-experienced raters *(Mdn)***	**Medium-experienced raters *(Mdn)***	**High-experienced raters *(Mdn)***
Content Positive	7.28	.03	0.00	0.00	0.00
Grammar Use Negative	9.57	.01	19.16	13.33	10.83
Intelligibility Positive	8.65	.01	0.00	0.00	0.00
L1 Use Negative	6.32	.04	0.00	0.00	0.00
Relevance Positive	7.10	.03	0.00	0.00	0.00

As can be seen in Table 66, there were statistically significant differences in "Content Positive," "Grammar Use Negative," Intelligibility Positive," "L1 Use Negative," and "Relevance Positive." Low-experienced raters significantly provided more "Grammar Use Negative" reasons (*Mdn* = 19.16%) than medium-experienced (*Mdn* = 13.33%) and high-experienced raters [(*Mdn* = 10.83%), *H* (2, *n* = 25) = 9.57, *p* = .01]. Thus, it was clear that high-experienced raters tended to give less negative than medium-, and low-experienced raters.

Follow-up Mann-Whitney *U* tests were conducted to compare the figures between the three rater experience groups for the positive and negative connotations of written score explanations. The statistically significant differences that the Mann-Whitney *U* tests revealed are summarized for each speaking performance pair in Table 67.

Table 67

Mann-Whitney *U* test results for the positive and negative written score explanations across rater experience groups

Positive and Negative Score Explanations	**Rater Groups**	***n***	***Mdn***	***U***	***z***	***p***	***r***
Content Positive	Medium	7	0.00	189.0	-2.42	.02	.62
	High	8	0.00				
Grammar Use Negative	Low	10	19.16	191.5	-2.94	.003	.69
	High	8	10.83				
Intelligibility Positive	Low	10	0.00	264.0	-2.70	.01	.63
	High	8	0.00				
L1 Use Negative	Low	10	0.00	276.0	-2.50	.01	.61
	High	8	0.00				
Relevance Positive	Medium	7	1.67	139.0	-2.61	.01	.67
	High	8	5.00				
Relevance Positive	Low	10	0.00	249.0	-1.98	.047	.46
	High	8	5.00				

According to the findings retrieved from Mann Whitney *U* tests, it was revealed that low-experienced raters statistically used more "Grammar Use Negative" (*Mdn* = 19.16%) than high-experienced raters [(*Mdn* = 10.83%), *U* = 191.5, *z* = -2.94, *p* = .003, *r* = .69]. However, there were not any statistically significant differences for "Grammar Use Negative" across low-experienced and medium-experienced; high-experienced, and medium-experienced raters. Another important finding to mention was that high-

experienced raters statistically provided more "Relevance Positive" (Mdn = 5.00%) than medium-experienced raters [(*Mdn* = 1.67%), *U* = 139.0, *z* = -2.61, *p* = .01, *r* = .67]. Similarly, high-experienced raters used more "Relevance Positive" (Mdn = 5.00%) than low-experienced raters [(*Mdn* = 0.00%), *U* = 249.0, *z* = -1.98, *p* = .047, *r* = .46].

4.5. Summary of the Results and Findings

This part of the dissertation provides information on the summary of findings based on each research questions. Tables 68 and 69 show the statistical findings of pertaining to RQ1 and RQ2. Furthermore, Tables 70 and 71 illustrate the findings related to G-study framework. Finally, Tables 72 and 73 summarize the results related to the qualitative findings.

Table 68

Summary of the results for RQ1

Research Question 1	**Result**
Are there any significant differences among the analytic scores of low-, medium- and high- quality L2 speaking performances?	Yes. The analytic scores assigned to low-quality, medium-quality, and high-quality L2 speaking performances showed statistically significant differences from each other.

As for RQ1, there were significantly different scores across low-, medium-, and high-quality L2 speaking performances. That is to say, the scores that the raters awarded differed across three speaking performance qualities.

Table 69

Summary of the results for RQ2

Research Question 2	**Result**
Are there any significant differences among the analytic scores assigned by low-, medium- and high experienced raters?	No. There were not any statistical differences in the total scores assigned to low-quality, medium-quality, and high-quality L2 speaking performances by low-experienced, medium-experienced, and high-experienced raters. In addition, no significant differences were revealed for the rubric component scores for all speaking performance qualities by three rater groups.
	As for examining the scores assigned to each speaking performance, there was only one statistically significant difference in a score assigned to a medium-quality response (SP29). This difference was only between medium-experienced and high-experienced raters.

The findings for RQ2 showed that significant differences among three rater experience groups were not observed in the ratings of low-, medium-, and high-quality L2 speaking performances. Additionally, there were not any statistically significant differences as for the scores assigned to components in the rubric. Specifically, the medium-quality speaking performance denoted by SP29 showed a statistical variation between medium-, and high-experienced raters.

Table 70 summarizes the findings of third research question. In this G-study, both cumulative and individual G-study analyses were carried out to predict source of variance and their effects on the assigned scores by speaking performance qualities. Following this, various D-studies were conducted for measurement designs pertaining to speaking performance qualities.

Table 70

Summary of the results for RQ3

Research Question 3	**Result**
What are the sources of score variation that contribute most (relatively) to the score variability of the analytic scores of L2 speaking performances?	In the person-by-rater-quality *(p x r x q)* crossed design, the person (*p*), namely students, was the largest source of variance (49.9%). This figure was followed by the residual variance (28.1%), which was the interaction across students, rater groups, speaking performance quality, and other unknown sources. The third largest variance was the interplay between persons and raters (12.7%). The interaction between raters and speaking performance quality had also somewhat effect on the variance (9.3%). Finally, no source of variation stemmed from raters (*r*), speaking performance quality (*q*), and the interaction between persons and quality.
	In the person-by-rater-quality *(p x r)* crossed designs, the residual component was the largest source of variance for low-quality (40.9%), medium-quality (51.9%), and high-quality L2 speaking performances (60.2%). Given the inconsistency of raters, the rater (*r*) variance was bigger for low-quality (31.7%) than medium-quality (19.6%), and high-quality L2 speaking performances (13.3%). As for the variability source stemming from students (*p*), the figures showed similarities across low-quality (27.4%), medium-quality (28.5%), and high-quality L2 speaking performances (26.5%), all of which illustrated the level of homogeneity of assigned scores.

Table 71 illustrates the generalizability and reliability coefficients depending on the speaking performance qualities and rater experience groups. The main aim of this analysis was to observe the rank of consistency across rater experience groups. Furthermore, D-studies were conducted to reveal the optimal number of raters in an ideal measurement design.

Table 71

Summary of the results for RQ4

Research Question 4	**Result**
Does the reliability (*e.g.,* dependability coefficients for criterion-referenced score interpretations) of the analytic scores of raters (low, medium and high) differ from each other?	In the person-by-rater-quality *(p x r x q)* crossed design, high generalizability and dependability coefficients were recorded for all speaking performance qualities across all rater groups.
	While grading low-quality L2 speaking performances, low-experienced raters (Ep^2 = .88 Φ = .80). were considerably more consistent than medium-, and high-experienced raters.
	As for medium-quality L2 speaking performances, medium-experienced raters rated more consistently than the other two rater experience groups. Specifically, high-experienced raters had the lowest coefficient indices (Ep^2 = .78 Φ = .73).
	Given the rating of high-quality L2 speaking performances, high-experienced (Φ = .84) and medium-experienced raters (Φ = .78) were relatively more consistent. However, medium-experienced raters showed considerable variation with the lowest coefficient indices (Ep^2 = .76 Φ = .69).

Table 72 shows the major findings retrieved from verbal protocols and written score explanations as regards to speaking performance qualities. However, Table 72 illustrates the decision-making behaviors and aspects of L2 speaking performances that rater experience groups paid attention to.

Table 72

Summary of the findings for RQ5

Research Question 5	**Findings**
How do raters make decisions while rating varying quality L2 speaking performances analytically?	Raters used more judgment strategies than interpretation strategies across all speaking performance qualities. In addition, raters utilized more language-focused strategies than self-monitoring and rhetorical strategies. However, these differences were not statistically significant. As for self-monitoring strategies, there were two statistically significant findings: a) considering personal situation of the test takers b) stating the overall performance of the test takers. 'Evaluating topic development' was the only statistically significant difference in all rhetorical and ideational focus strategies. 'Evaluating intelligibility of the response' and 'Evaluating fluency' were statistically used more frequently than other language focus strategies. All in all, 'Evaluating fluency', Evaluating task completion', and 'Stating or revisiting scoring' were the most commonly used strategies for all speaking performance qualities. Fluency, Vocabulary, and Grammar Use were the top three commonly used written score explanations across all speaking performance qualities.

Table 73

Summary of the findings for RQ6

Research Question 6	**Findings**
How does professional experience affect raters' decision-making processes and the aspects of speaking responses they focus on?	Medium-experienced raters statistically used more 'Self-monitoring' strategies than low-experienced and high-experienced raters. Similarly, medium-experienced raters statistically employed more 'Self-monitoring judgment strategies' than low-experienced and high-experienced raters. As for self-monitoring strategies, medium-experienced raters statistically used the strategy called 'Stating or revisiting scoring' more frequently than low-, and high-experienced raters. 'Interpreting vague or equivocal expressions' and 'Evaluating topic development' were two rhetorical and ideational focus strategies that rater groups statistically differed. With regard to language focus strategies, medium-experienced and high-experienced raters showed statistically significant differences while using the strategy called 'Rephrasing responses for interpretation'. Overall, 'Evaluating task completion', 'Stating or revisiting scoring' and 'Evaluating fluency' were top three commonly used decision-making behaviors by low-experienced, medium-experienced, and high-experienced raters while grading all L2 speaking performances. Rater experience groups statistically showed differences while using the written score explanations called 'Grammar Use', 'Vocabulary', Intelligibility', 'L1 Use', and 'Relevance'.

The findings pertaining to both quantitative and qualitative data presented in this summary section are explained in considerable detail in the next chapter.

CHAPTER V

Discussion and Conclusion

5.1. Introduction

In this chapter, the findings from the previous chapter are discussed within each related question as regards to six main areas: a) differences across speaking performance qualities, b) differences among rater experience groups, c) the sources of score variation contributing to the score variation of speaking performance qualities, d) the reliability of the analytic scores of the raters, e) raters' decision making behaviors while rating different quality L2 speaking performances, and f) the effect of professional experience on raters' decision-making processes and the aspects they focus on. The following sections include limitation of the study, conclusion, pedagogical and methodological implications, and future research.

5.2. Speaking performance qualities and raters experience groups (RQ1 and RQ2)

The L2 speaking performances, classified as low-, medium-, and high-quality for this study, were collected from an EPP at a technical university in western Türkiye. All participants contributing to this study were employed as full-time instructors at this university, and were EFL professionals graduating from the departments of language teaching such as ELL, and ELT. During the rater training sessions, all raters were informed about the language profile of the students and the level of the speaking test from which the L2 speaking performances were collected. However, they were not especially informed about the quality division of each speaking performance.

The findings of RQ1 showed that significantly different scores were assigned for low-, medium-, and high-quality L2 speaking performances by raters. That is to say, the scores that the raters awarded were distinctive across three speaking performance qualities.

In fact, even if the speaking performance quality division information was not provided, the raters were mostly successful in recognizing the varying speaking performance qualities in this study. Since the relationship across essay quality, assigned scores, and rating patterns is one of the under-researched areas in L2 speaking assessment, it would be useful to review the results and findings of L2 writing assessment. (Brown, 1991; Daly & Dickson-Markman, 1982; Engber, 1995; Freedman, 1981; Han, 2017; Şahan, 2019). The common point of all these studies was that text quality could interact with numerous factors such as assessment context, the comparison with previously assigned scores, the order of essay qualities, rating training, and the perception of text quality. Thus, it is obvious that different quality L2 speaking performances or essays might cause fairness and reliability issues in L2 performance assessment. While raters in this study could generally determine the right group of speaking performance quality, there was some confusion as to the quality distinction of high-quality performances.

Looking at the dispersion of speaking performance qualities in detail, I can say that there was more variation of assigned scores for high-quality L2 speaking performances than low-, and medium-quality L2 speaking performances. This finding showed some parallelisms with the findings of an L2 writing assessment based study (Şahan, 2019), where the researcher revealed more variation across the scores given to high-quality essays. However, another study that investigated essay quality (Han, 2017) found that there was more variation in the scores of low-, and medium quality essays than high-quality ones. Even if the raters contributing to this study received rating trainings and were familiar with the institutional assessment procedures, there was somewhat higher variation for high-quality L2 speaking performances. The reason of this variation could be related to raters' self-theories, personal assumptions, error frequency, and (un)familiarity with the students or examinations (Daly & Dickson-Markman, 1982; Engber, 1995; Freedman, 1981). For instance, Engber (1995) found that the more errors essays had, the lower scores raters tended to assign. Supposing that there were relatively more errors in the low-, and medium-quality L2 speaking performances than high-quality ones, raters might have scored weaker speaking performances more consistently.

As for rater experience groups, low-experienced and medium-experienced raters tended to assign higher scores to three speaking performance qualities than high-experienced ones did. This also means that high-experienced raters were slightly stricter while less experienced raters were relatively more lenient. Although low-experienced raters seemed to be more lenient than experienced raters, the qualitative data analysis revealed that medium-experienced and high-experienced raters expressed more positive comments (40% and 39.16%, respectively) than low-experienced raters (22.49%). However, the total and rubric component scores that low-, medium-, and high-experienced raters assigned did not show considerable differences from each other while rating varying quality L2 speaking performances. This finding might be related to the rater experience scale form that I utilized while categorizing the experience group of raters. Thanks to this experience scale, I was able to determine the participants' actual rating experience by minimizing the effect of personal beliefs and experience gained from other institutions. For instance, a participant with 10 years teaching and assessment experience was classified as a low-experienced rater for this study because this rater had very little speaking assessment experience at the institution where this study was conducted. Supporting the quantity of rating in actual experience, previous research revealed that inexperienced raters' steady progress in rating consistency could be made through regular practice (Lim, 2011). Therefore, in this study, I prioritized the bulk of institutional speaking assessment experience instead of years of experience in general.

Similar to the situation of research on text quality, it would not be wrong to express that the number of rater's rating experience within the L2 writing assessment context outweighs the ones in L2 speaking assessment (Barkaoui, 2010a; Cumming 1990; Delaruelle, 1997; Leckie & Baird, 2011; Lim, 2011; Myford et al., 1996; Sakyi, 2003; Song & Caruso, 1996; Wolfe et al., 1998). While some of these studies found that experienced raters were more homogenous and consistent, some of them investigated the effect of added practice on rating performance, and others observed no differences between experienced and inexperienced raters. Given rater experience research in speaking assessment, Davis (2016) found that low-experienced speaking raters, who were selected from high-experienced teachers, made slight progress after rater training sessions. In another study, Bonk and Ockey (2003) found that less experienced speaking raters were more inconsistent than the

experienced ones. However, the raters whose scores were not reliable improved their rating performance following rater training sessions and regular practice. Examining three rater experience groups, Kim (2015) reported that experienced raters had less inconsistencies and variations than novice and developing raters, who made considerable progress in rating speaking performances after a couple of rater training sessions. Even though this research did not observe the effect of rater training and rater experience on score variation as the aforementioned speaking assessment studies did, raters with three different rater experience groups was the central focus in this study. High reliability coefficients were recorded for low-experienced raters (.95), medium-experienced raters (.94), and high-experienced raters (.94) while scoring all speaking performance qualities (low-, medium-, and high-quality). Relatively lower coefficients were, nevertheless, revealed for all rater experience groups while rating each speaking performance. For instance, low-experienced raters (.80) were considerably more consistent than medium-, and high-experienced raters (.68 and .70) while grading low-quality L2 speaking performances. Grading medium-quality L2 speaking performances, medium-experienced raters and low experienced raters showed similar reliability coefficients (.80 and .78), but high-experienced raters had the lowest coefficient (.73). Similar coefficients were obtained for the raters while assessing high-quality L2 speaking performances (.78, .69, .84, respectively), showing that high-experienced raters were more consistent than the other two experience groups. Raters' inconsistencies within each experience group might be related to the areas where raters confused while interpreting rubric descriptors. Although there were some lower coefficient figures, rater groups (low-, medium-, and high-experienced) seemed to show a similar tendency across speaking performance qualities. The reason of this finding might be related to the rater experience scale, through which participating raters were selected and grouped. This methodology was also used in the studies (Davis, 2016; Kim, 2015), in which the researchers relied on a set of criterion before deciding the quality group of raters. Most importantly, the fact that raters were formed from the same institution might have been the reason of relatively homogenous variation across rater experience groups.

In conclusion, the ratings of L2 speaking performances interacted with both the speaking performance qualities and rater experience groups. Initially, the scores that the raters assigned while rating low-, medium-, and high-quality L2 speaking performances

showed marked differences from each other, implying that the rater groups were able to spot the speaking performance qualities. At the same time, there was slightly more variation within the scores of high-quality L2 speaking performances compared to low-, and medium-qualities. In the second place, although there were not any statistically significant differences in the total and component scores that rater experience groups assigned, high-experienced raters tended to award lower scores than low-, and medium-experienced raters. Finally, the rater reliability coefficients illustrated that low-experienced raters were consistent while grading low-quality L2 speaking performances; medium-, and low-experienced raters were somewhat more reliable than high-experienced raters while rating medium-quality L2 speaking performances; high-experienced raters were more consistent than low-, and medium-experienced raters while assessing high-quality L2 speaking performances. Despite these slight differences, the coefficients within each rater group across three L2 speaking performances did not show a widening gap. As can be seen, numerous factors might be in a dynamic relationship with both speaking performance quality and rater experience. Therefore, these findings suggest that forming experience groups based on a set of criterion may yield balanced results while creating a consistent speaking rater model.

5.3. Generalizability and dependability coefficients for speaking performance qualities and rater experience groups (RQ3 and RQ4)

The generalizability analysis results illustrated that persons (p) was the biggest source of variance (45.9%) as regards to the person-by-rater-by-quality (p *x* r *x* q) design. This result was expected since the speaking tasks in the test seemed to differentiate the performances of the test takers. Nevertheless, lower figures were obtained for the persons facet in terms of low-quality, medium-quality, and high-quality L2 speaking performances. This situation was also being expected because the person-by-rater (p *x* r) design was comprised of homogeneous student groups compared to the all L2 speaking performances within the same design. The persons facet for individual designs (p *x* r) showed similarities: low-quality L2 speaking performances (27.4%), medium-quality L2 speaking performances (28.5%), and high-quality L2 speaking performances (26.5%). Much lower figures of persons facet could have been obtained for individual designs. However, the quality division

raters might have confused some of the L2 speaking performances' quality during the division process.

Given all speaking performance qualities, namely (p *x* r *x* q) design, residual facet, which might be affected by raters, speaking performance quality, and other unexplained sources of errors, was the second largest source of variance (28.1%). The residual component was relatively higher in the individual designs of low-quality (40.9%), medium-quality (51.9%), and high-quality L2 speaking performances (60.2%) as there were less number of facets in the measurement design. These results might mean that score variation could have stemmed from other and unexplained factors such as scoring scales, personal beliefs, raters' training background (Bachman, 2004; Brennan, 1992, 2001, 2011). The findings in this study echo the results of a study (Kim, 2009a) where the researcher found higher dependability coefficients of persons (*p*) and residual variance components of NS and NNS speaking raters in the crossed G-theory design. In addition, Kim's study revealed that there was very little effect of rater facet (*r*) for both rater groups.

As for (p *x* r *x* q) design, there was no source of variation stemming from rater facet (*r*) (0%), showing that rater groups displayed a high-level of consistency while grading all L2 speaking performances in a cumulative manner. Once L2 speaking performances were examined individually, there were more issues of leniency and severity as regards to rater facet while assessing low-quality (31.7%), medium-quality (19.6%), and high-quality L2 speaking performances (13.3%). These findings may imply that raters had more difficulties in grading individual quality of L2 speaking performances than mixed quality L2 speaking performances. In this study, it seems that raters were more inconsistent while assessing L2 speaking performances with weaker performance than medium-, and high-quality L2 speaking performances.

The interplay between persons (*p*) and raters (*r*) (12.7%) was the third largest source of variance in all L2 speaking performances design. This amount of variance suggests that there were some inconsistencies in some scores that a group of raters awarded. Following that, the component between raters (*r*) and speaking performance quality (*q*) (9.3%) was the

fourth biggest variance, illustrating that some of the raters showed variance while grading some individual speaking performance qualities. The other source of variance components such as speaking performance quality (*q*) (0%), and the interaction between persons and speaking performance quality (0%) did not seem to have any impacts on the score variation.

Considering mixed (p *x* r *x* q) and individual generalizability (p *x* r) designs, high dependability coefficients were obtained for all speaking performance qualities (.98) and low-quality (.90), medium-quality (.91), and high-quality (.90) across three rater experience groups. The fixed number of facets, especially the larger number of rater facets, might have contributed to these high dependability figures (Brennan, 2001; Shavelson & Webb, 1991). Therefore, it would be useful to examine the facets when the numbers were decreased (Taşdelen Teker & Güler, 2019). Given that the ideal reliability coefficient should above .80 and the total of number of raters is 25 in this study, for all L2 speaking performances (p *x* r *x* q), three raters would be sufficient to provide the reliability of assigned scores. As for the scenario for the low-quality L2 speaking performances, a number of 11 raters would provide an acceptable level of dependability coefficient. Similarly, 11 raters would still give reliable results for medium-quality L2 speaking performances. Lastly, a number of 12 raters would be adequate to sustain consistency of ratings for high-quality L2 speaking performances. The findings retrieved from the all L2 speaking performances design might work for high-stake speaking examinations since the language level of the students are likely to be mixed. Thus, a number of 3 raters from the (p *x* r *x* q) design would seem to be efficient. However, for speaking exams that aim to assess individual qualities at least 11 raters should be in the rating process. Although it is a study (Şahan, 2019) in which the researcher explored the dependability coefficients of EFL raters who assessed varying quality essays, the findings show parallelisms with this study. The researcher also found that three raters would be sufficient while grading mixed quality of EFL essays. In contradiction to the findings by Şahan (2019), in this study, rater groups drew a more consistent profile while assessing low-quality, medium-quality, and high-quality L2 speaking performances. This finding signifies the importance of handling all qualities and individual qualities separately. However, researching the relationship between speaking test tasks and other facets would be necessary to minimize reliability threats. For instance, the findings of a study (Bachman et. al., 1995) in which the researcher investigated the tasks, scores, and raters in an immersion speaking

test by means of G-theory and Many-facet Rasch measurement revealed that some of the certain raters and tasks decreased the reliability of assigned scores. Similarly, researching the impact of tasks and ratings in the TOEFL speaking exam, Lee (2005) found that increasing the number of speaking tasks rather than the ratings provided the highest dependability coefficients, suggesting an alternative framework for increasing reliability. Therefore, the interaction between individual spoken qualities, speaking exam tasks, raters, and other unexplained factors needs to considered in detail while designing a fair and reliable rating system in L2 speaking examinations.

To sum up, the speaking tasks were able to separate the distinct student performances within all qualities of L2 speaking performances. As for the separate analysis of low-, medium-, and high speaking performance qualities, the variation across students showed similarities. Thirdly, score fluctuations stemming from residual component were noticeable in all mixed and individual L2 speaking performances. This might refer to the effect of systematic or random source of errors on score variation. In addition, larger variations of residual were observed for individual L2 speaking performances than the mixed qualities. Finally, the analysis of interaction between raters and speaking performance quality illustrated that raters had more issues of consistency in terms of rating low-quality L2 speaking performances than medium-, and high-quality L2 speaking performances.

5.4. Raters' decision-making behaviors within the scope of speaking performance qualities and rater experience groups (RQ5 and RQ6)

As for qualitative data**,** verbal protocols and written score explanations were utilized to investigate the effect of speaking performance quality and rater experience on raters' decision-making behaviors while grading low-, medium-, and high-quality L2 speaking performances. When the strategies were considered collectively, more judgment strategies were employed than interpretation strategies across all speaking performance qualities. This finding showed parallelisms with some of the studies that investigated writing raters' decision-making behaviors (Barkoui, 2010b, Cumming et. al. 2002, Gebril & Plakans, 2014). However, in another study, Şahan and Razı (2020) found that more interpretation

strategies were used more than judgment strategies. Above all, this might be related to the fundamental differences between L2 speaking and writing assessment nature, the former of which focuses on the flowing data set and the latter of which is engaged with the written data. It would be useful to note here that there is scarcity of studies investigating speaking rater's decision-making behaviors. Secondly, certain factors such as scoring rubrics, training history, raters' background and perceptions might have contributed to this finding. The descriptors in the scoring rubric used for this study might have also affected the type of strategies that the raters employed. That is to say, raters might have evaluated judgment strategies such as fluency, vocabulary, grammar use, sentence types, topic development, and task completion to correspond with the scoring rubric that they used.

High-quality L2 speaking performances elicited more self-monitoring focused strategies. However, raters used more rhetorical focused strategies while grading medium-quality L2 speaking performances. At the same time, raters relied on using language focused strategies more frequently while evaluating low-quality L2 speaking performances. Given the slight differences across speaking performance qualities, none of these figures were statistically significant. This finding might be related to the context of assessment in which the participating raters had already been familiar with the students, speaking test, and scoring rubric.

Looking at the individual strategies by speaking performance qualities, there were three focus areas in the coding scheme: a) self-monitoring focus strategies, b) rhetorical and ideational focus strategies, and c) language focus strategies. Firstly, two self-monitoring focused strategies were used more significantly: a) raters mostly considered personal situation of the test takers while rating low-quality L2 speaking performances, and b) raters stated the overall performance of test takers while grading high-quality L2 speaking performances. Secondly, raters significantly relied on only one rhetorical and ideational focused strategy in which they mostly evaluated students' topic development in scoring medium-quality and high-quality L2 speaking performances. Thirdly, two language focused strategies were significantly salient: a) raters largely evaluated the intelligibility of the response in handling low-quality L2 speaking performances, and b) raters mostly tended to evaluate fluency of the weaker students' performances. These differences across speaking

performance qualities might be related to the competence that each proficiency requires (Cumming et al., 2002). To illustrate, raters tended to pay attention to students' topic development while grading higher proficiency level performances. Overall, the top three strategies that raters preferred showed similarities across low-, medium-, and high-quality L2 speaking performances: a) state or revisit scoring, b) evaluate task completion, content, and relevance, and c) evaluate fluency. With regard to the top ten most frequently used strategies, all speaking performance qualities elicited more language focused strategies than self-monitoring and rhetorical strategies. In fact, raters particularly became engaged in forms rather than content and discourse of responses. This finding echoes the results of a study conducted by Han (2017) in which the researcher found that all essay qualities attracted language focused strategies. Having mentioned before, I can relate these differences with the expectations stemming from scoring rubric, institutional objectives and outcomes, and raters' background. In addition, raters might have put interpretations on some of the scale descriptors that they found incomplete or weak (Orr, 2002).

The top five most frequently used written score explanations were fluency, vocabulary, grammar use, task completion, and topic development for all speaking performance qualities. Corroborating the findings retrieved from verbal protocols, the majority of the written score explanations were language oriented reasoning such as fluency, vocabulary, and grammar use. The score explanation "pronunciation" was significantly elicited more for low-quality L2 speaking performances. Except for "pronunciation, there were not, however, any other significant differences in score explanations across low-, medium-, and high-quality L2 speaking performances. Similar to this finding, some of the studies exploring speaking rater's cognition (e.g., Brown et. al., 2005; Cai, 2015; Chalhoub-Deville, 1995) revealed that pronunciation was a salient point in raters' decision-making behaviors. As regards to the positive and negative types of written score explanations, raters significantly made more negative comments on low-quality L2 speaking performances than medium-, and high-qualities. This finding showed similarities with the results of a study (Barkaoui, 2010a) where the researcher revealed positive reasons for high-quality essays and negative explanations for low-quality essays. This finding might corroborate the quantitative results from RQ1 in which raters tended to give lower scores to low-quality responses,

medium scores to medium-quality responses, and higher scores to high-quality L2 speaking performances.

When I examined the major categories of decision-making behaviors by rater experience, I found that all rater groups reported more judgment strategies than interpretation strategies across all speaking performance qualities. At the same time, language focus was the most frequently used strategy type of all foci, showing certain similarities with the findings of Cai (2015). High-experienced raters used language-focused strategies slightly more than low-, and medium-experienced raters. However, none of these differences across rater groups were statistically significant. Medium-experienced raters significantly used more self-monitoring focused strategies than low-experienced and high-experienced raters. Furthermore, medium-experienced raters significantly relied on self-monitoring focused judgment strategies more than low-, and high-experienced raters. Overall, medium-experienced raters significantly separated from low-, and high-experienced raters in terms of self-monitoring and self-monitoring judgment strategies.

With regard to individual strategies, first of all, medium-experienced raters significantly stated or revisited their scoring more than low-experienced and high-experienced raters. This strategy was within the category of self-monitoring strategies. Secondly, when I analyzed rhetorical and ideational focus strategies, I found that low-experienced raters significantly interpreted vague expressions more than medium-, and high-experienced raters. In addition, high-experienced raters significantly evaluated students' topic development more than low-experienced and medium-experienced raters. Thirdly, as for the language focused strategies, there was only one significant finding. High-experienced raters mostly rephrased responses for interpretation more than medium-experienced raters. These variations might give us implications about the rating patterns that each experience group followed. Given the top five most frequently used strategies by rater groups across all speaking performance qualities, I observed certain similarities across rater experience groups (e.g., evaluate fluency, vocabulary, task completion and content).

Comparing each rater group's decision-making behaviors across speaking performance qualities, I revealed that only high-experienced raters showed statistically significant differences in two categories: a) rhetorical focus strategies, and b) language focus judgment strategies. High-experienced raters significantly used more rhetorical strategies while grading medium-quality L2 speaking performances. Additionally, high-experienced raters significantly paid attention to language focused strategies while scoring low-quality and high-quality L2 speaking performances. These findings point to a close relationship between rater experience and speaking performance quality. Although all raters were from the same institution and used the same scoring rubric, only high-experienced raters displayed statistically significant differences across response qualities. This complexity might be related to the rating approaches that each specific rater experience group adopted (Ang-Aw & Goh, 2011; Pollitt & Murray, 1996).

In terms of written score explanations by rater experience groups, there were statistically significant differences in five categories: grammar use, vocabulary, intelligibility, L1 use, and relevance. For instance, low-experienced raters significantly reported "grammar use" more than high-experienced raters. That is to say, higher experienced raters might have felt relatively more confident about grammatical aspects while determining the level of performance. However, medium-experienced and high-experienced raters significantly attended to "vocabulary" more than low-experienced raters. This finding might mean that the aspects that raters paid attention to could be a distinguishing feature for rater's decision-making behaviors.

To conclude, speaking performance quality and rater experience seemed to have an impact on the rating behaviors. Collectively, raters used more judgment strategies than interpretation strategies across all speaking performance qualities. In addition to this, each speaking performance elicited a different strategy area. For instance, the reason why low-quality L2 speaking performances attracted language focus strategies more than others could be related to their limited content and relevance. When I examined the individual areas of strategies, it was clear that the level of speaking performance quality determined the type of strategy used. Another overall result was that all speaking performance qualities elicited more language focused strategies than self-monitoring and rhetorical strategies. As for the

major categories of strategies, only medium-experienced raters showed statistically significant differences across all rater groups. When I examined each rater experience group within speaking performance qualities, I found that only high-experienced raters significantly used more rhetorical strategies and language focused strategies. All in all, the endeavor to understand complexities of rater's thinking process stemming from speaking performance qualities and rater experience might be beneficial to L2 speaking assessment contexts in terms of rater training, speaking task design, and rubric development.

5.5. Limitations of the Study

The raters in this study had periodically attended rater training sessions organized by the professional development unit of the department as well as calibration meetings before each final speaking exam. Furthermore, participating raters were guided through orientation and tutorial sessions to boost familiarity with the analytic scale prior to actual rating. However, the lack of professional rater training might have had an impact on the score variation in this study. In fact, many studies found that long-term rater training sessions that aim to track speaking rater's rating development through feedback channels improved the reliability of assigned scores (Papajohn, 2002; Stitt et. al., 2003; Wigglesworth, 1993; Xi & Mollaun, 2011, Yan, 2014).

The topics and tasks used in speaking exams can affect numerous aspects of assessment process since this area consists of crucial issues such as task difficulty, gender bias, background knowledge, and the type of tasks (Fulcher & Reiter, 2003; Huang et al., 2018; Teng, 2007; Khabbazbashi, 2016; Lumley & O'Sullivan, 2005; Tavakoli, 2009; Weir & Wu, 2006). Although the main data were collected from this institutional speaking final exam to eliminate certain disadvantages stemming from task related issues, various results might have been obtained if the task and topic conditions had been controlled. This definitely opens up new avenues for researching the effect of speaking tasks and topics on score variation and raters' rating behaviors.

Previous research investigated the advantages and disadvantages of using analytic and holistic scales in L2 speaking assessment (Brown, 2007; Brown, 2006; Chuang, 2009; Fulcher et. al., 2010; Isaacs & Thomson, 2013; Upshur & Turner, 1995). In this study, given the certain advantages of analytic scales for speaking tests, I utilized the institutional analytic speaking scale that the raters were already familiar with. However, the use of an alternative scale might have affected the scores and rater's thinking tendencies.

The use of verbal protocols was one of the serious challenges that the raters faced during the rating process. Four measures were adopted to minimize potential problems: a) a clear guideline, b) a group session for verbal protocols, c) one-to-one feedback on a regular basis, and d) a sample tutorial video. In addition, the raters were allowed to complete verbal protocols at home. Otherwise, they could have felt themselves restricted or been put under pressure (Barkaoui, 2010a; Cumming et. al. 2002). Another limitation was that the raters were not restricted, and thus they were free to complete verbal protocols at home within the expected deadline. Considering all these points, the results of this study would have been different if verbal protocols had been collected in a controlled research design.

Finally, high variances of residual component were obtained for all L2 speaking performances (28.1%), low-quality L2 speaking performances (40.9%), medium-quality L2 speaking performances (51.9%), and high-quality L2 speaking performances (60.2%) in this study. Given that residual component is related to the other unknown systematic or unsystematic sources of variance, a lower figure of this component would have given different indices for the other facets such as persons, raters, and quality (Brennan, 2001; Shavelson & Webb, 1991). Besides this limitation, the collaboration of both G-theory and Many-facet Rasch measurement could have been needed to estimate the effect of each task and rater on score variability (Bachman et. al., 1995).

5.6. Conclusion

First, the inferential statistics analysis showed that the scores assigned to low-, medium-, and high-quality L2 speaking performances showed significant differences from

each other. That is to say, all raters were able to notice the differences in speaking performance quality groups although no explanation was made regarding the quality division of L2 speaking performances.

Second, rater experience groups did not show statistically significant differences in their total and rubric component scores. However, one of the scores assigned to a medium-quality response showed a statistically significant difference, which was between medium-experienced and high-experienced raters. Overall, there was not a considerable impact of rater experience on the assigned scores.

Third, the effect of rater facet as a source of variation was very limited when all L2 speaking performances were examined in total. However, the rater facet had a substantial impact on the variation of scores when low-, medium-, and high-quality L2 speaking performances were analyzed individually. This showed that rater experience groups seemed more inconsistent with separate L2 speaking performances than all mixed L2 speaking performances. Furthermore, the component of residual variance was relatively higher for both G-study designs consisting of all L2 speaking performances ($p \times r \times q$) and individual speaking performance ($p \times r$). Due to these high residual indices, the sources of variance that were not included in the designs such as rater training, speaking tasks and topics, rater's education background might have been hidden.

Fourth, high dependability coefficient indices were obtained for mixed quality and individual speaking performance qualities. Given the findings of D-studies, I found that three raters would still be reliable while grading mixed quality L2 speaking performances. When low-, medium-, and high-quality L2 speaking performances were analyzed separately, the results of D-studies showed that a number of 11 to 12 raters would award scores consistently.

Finally, both speaking performance qualities and rater experience groups elicited various decision-making behaviors. Considering general categories, raters used more judgment strategies than interpretation strategies across all speaking performance qualities. However, in terms of strategy focus types, raters paid attention to more language focused

strategies than self-monitoring and rhetorical focus strategies. Similar to this, raters tended to produce language related written score explanations. Given the results for rater experience groups, medium-experienced raters significantly separated from low-, and high-experienced raters regarding self-monitoring and self-monitoring judgment strategies. When I analyzed each rater experience group's strategy use within each speaking performance qualities, I revealed that only high experienced raters showed statistically significant differences in rhetorical focus and language focus judgment strategies.

5.7. Practical Implications

This study has several practical implications for rater training, the regulation of speaking rater protocols, and speaking test rubrics. Firstly, the results highlight the importance of comprehensive and regular rater training sessions for raters. The institution where the main data of this study was collected from has been implementing a well-established assessment system since they successfully completed two internationally recognized accreditation processes. Within this perspective, the institution provides calibration meetings and rater training sessions before end-quarter examinations. In addition, all these advancements in assessment have been regulated by the administration and teaching directorate. However, the rater trainings organized in this institution were not detailed and tailor-made to suit the rater's rating needs and lacks. Therefore, an effective rater training model presenting authentic decision-making behaviors and providing guided feedback would improve the reliability of assigned scores.

The management of rating quality is one of the crucial aspects of speaking assessment process. A rater experience scale form had not been available until this study was conducted at this institution. With the help of this scale, raters were placed in the right experience group to the extent that they had in-house speaking assessment experience. The results of this study showed that rater experience groups showed similarities in certain areas while rating different speaking performance qualities. That is why, determining the actual experience of speaking raters can minimize the threats to score reliability. Furthermore, pairing raters from varying experiences would improve the chance of success in exchanging

different rating perspectives. Otherwise, rater's self-described rater experience or rating experience coming from out of in-house context would not reflect rater's real experience background. Therefore, an inventory of rater's experience can be updated on a regular basis using a rater experience scale form. This might also contribute to determining rater's professional development needs.

Given that there is no a perfect scoring rubric, the use of either holistic or analytic rating scales might pose challenges or provide solutions to assessment issues (Brown, 2007; Brown, 2006; Chalhoub-Deville, 1995; Chuang, 2009; Fulcher et al., 2010; Green, 2014; Isaacs & Thomson, 2013; Madsen, 1983; Upshur & Turner, 1995). Instead of relying on one type of scoring scale, a step-by-step guide can be formed so that speaking raters might interact with the descriptors in the scale. It is crucial to note here that institutions can prepare user-friendly and clear guidelines by elaborating on commonly used decision-making strategies that can be retrieved from institutional in-house speaking examinations.

5.8. Methodological Implications

This study has several implications for methodology. First, the study explored the impact of speaking performance quality and rater's experience on score variation by means of an institutional analytic rubric. Speaking raters might not be able to use speaking rubrics effectively due to factors such as instant nature of speaking exams, limited time, accents, and gestures (Winke, 2012). The comparison of various rubric types might therefore produce better results in terms of mapping how raters give decisions while scoring different quality L2 speaking performances (Kim, 2006; Luoma, 2004).

In addition, the scores that were used in this study were collected from a real exam context. The rationale behind this decision was that students would take the process seriously and raters would score the performances that they were familiar with. Alternatively, the data collected from an out of exam context can be implemented to observe students' performances in a less stressful environment. Additionally, a set of speaking exam data can be collected via a computer program, by means of which students can record their

performance by themselves. This might reduce the effects of factors such as examiners and anxiety on speaking assessment process.

Considering the use of verbal protocols in this study, both theoretical and practical information were organized for raters via a sample training video, general and tutorial sessions. On demand, the researcher gave personal support and feedback to the raters. However, follow-up interviews on the completion of verbal protocols might provide a better understanding how raters conducted the verbal protocols and score reasoning. For instance, the researcher and rater can listen to the sub-sample of verbal protocol recordings simultaneously, and then the researcher may ask questions about salient points. Furthermore, this kind of verbal protocol approach might offer explanations for the incomplete or ambiguous comments.

Lastly, this study demonstrated the use of G-theory framework with the aim of revealing the source of score variation in terms of certain components such as raters, students, speaking performance quality, and residual. As the results of G-theory analyses showed, the residual indices, namely hidden sources of variance, were obtained higher than they could have been. Therefore, this finding underlines the importance of investigating various facets such as task difficulty, speaking topic, rating length, and rater fatigue. Moreover, the corroboration of the results obtained from G-theory and Rasch measurement designs can be implemented to determine the effect of each single facet on the reliability of assigned scores.

5.9. Suggestions for Future Research

The findings, limitations, and implications of this study suggest new avenues for future research. First, this study explored the raters with a non-native speaker of English background and working as instructors in an EFL context. The description of patterns that NES and NNES raters use might have valuable implications for rater's background and cognition (Carey et al., 2011; Huang & Jun, 2015; Huang et al., 2016; Kang et al., 2019; Kim, 2009b; Wei & Llosa, 2015; Zhang & Elder, 2011). Additionally, raters with different

professional backgrounds can give new insights into rating behaviors and score variation (Douglas & Myers, 2000; Lumley, 1998). Therefore, this study suggests that future research would design rater groups from different backgrounds such as NES, NNES, students, expert raters and other professionals. For instance, future research can include expert raters working as a professional rater for TOEFL or IELTS to compare the scores assigned by novice raters.

Secondly, in this study an analytic scale was preferred because they are known to provide a more comprehensive scaling range to raters and also ample information of test takers' strengths and weaknesses. In fact, analytic scales can provide larger scaling range to raters and reveal test takers' strengths and weaknesses although their subcategories with detailed descriptors might not be very practical for raters especially in speaking tests (Alderson et al., 1995; Brown, 2004; Luoma, 2004; Madsen, 1983). Considering the drawbacks of using an adapted one single type scoring scale (either holistic or analytic), future research may focus on the development of an authentic and tailor-made speaking rubric through a process in which all participating raters can contribute to each descriptor in the scale.

Another recommendation for future research can be related to the investigation of rater experience and score variation in different EFL contexts. One of the major reasons why I opted to focus on the context of EPPs was to contribute to the development of an L2 speaking assessment model within the framework of EPPs. Given that EPPs organize numerous low-stakes and high-stakes examinations, the fairness and reliability of assigned scores in these exams need to be taken seriously. Therefore, expanding on the methodology of this study, future research can initiate studies that explore factors affecting score variation and rater's cognition, the results of which might ensure the quality of L2 speaking assessment procedures in EPPs. Additionally, the future research can observe the improvement of speaking raters who have received rater training in the long-term. With the help of these studies, effective rater training models for raters and rater trainers can be developed.

REFERENCES

Alderson, C. J., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation* (1st ed.). Cambridge University Press.

Ang-Aw, H. T., & Chuen Meng Goh, C. (2011). Understanding discrepancies in rater judgement on National-Level oral examination tasks. *RELC Journal*, *42*(1), 31–51. https://doi.org/10.1177/0033688210390226

Bachman, L. F. (1990). *Fundamental considerations in language testing* (1st ed.). Oxford University Press.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, *19*(4), 453–476. https://doi.org/10.1191/0265532202lt240oa

Bachman, L. F. (2004). *Statistical analyses for language assessment book*. Cambridge University Press.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, *12*(2), 238–257. https://doi.org/10.1177/026553229501200206

Bachman, L., & Palmer, A. (2010). *Language assessment in practice* (1st ed.). Oxford University Press.

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

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

Barkaoui, K. (2010a). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly*, *44*(1), 31–57. https://doi.org/10.5054/tq.2010.214047

Barkaoui, K. (2010b). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, *7*(1), 54–74. https://doi.org/10.15434300903464418

Bejar, I. I. (1985). A preliminary study of raters for the test of spoken English. *ETS Research Report Series*, *1985*(1), 1–28. https://doi.org/10.1002/j.2330-8516.1985.tb00090.x

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice*, *31*(3), 2–9. https://doi.org/10.1111/j.1745-3992.2012.00238.x

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, *20*(1), 89–110. https://doi.org/10.1191/0265532203lt245oa

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement*, *24*(4), 339–353. https://doi.org/10.1177/01466210022031796

Brennan, R. L. (2001). *Generalizability theory*. Springer Publishing.

Brennan, R. L. (2005). Generalizability theory. *Educational Measurement: Issues and Practice*, *11*(4), 27–34. https://doi.org/10.1111/j.1745-3992.1992.tb00260.x

Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education*, *24*(1), 1–21. https://doi.org/10.1080/08957347.2011.532417

Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of School Psychology*, *52*(1), 13–35. https://doi.org/10.1016/j.jsp.2013.11.008

British Council. (2015). *The state of English in higher education in Turkey*. TEPAV. https://www.britishcouncil.org.tr/sites/default/files/he_baseline_study_book_web_-_son.pdf

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, *12*(1), 1–15. https://doi.org/10.1177/026553229501200101

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, *20*(1), 1–25. https://doi.org/10.1191/0265532203lt242oa

Brown, A. (2006). An examination of the rating process in the revised IELTS speaking test. In: P. McGovern, & S. Walsh (Eds.), *IELTS Research Report: Volume 6* (pp. 41–70). IELTS Australia & British Council. https://www.ielts.org/-/media/research-reports/ielts_rr_volume06_report2.ashx

Brown, A. (2007). An investigation of the rating process in the IELTS oral interview. In L. Taylor & P. Falvey (Eds.), *Research in speaking and writing assessment IELTS collected papers* (pp. 98–141). Cambridge University Press.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English for academic purposes speaking tasks*. ETS Research Report Series, 2005(1). https://doi.org/10.1002/j.2333-8504.2005.tb01982.x

Brown, D. H. (2004). *Language assessment - principles and classroom practice* (1st ed.). Pearson ESL.

Brown, G., & Yule, G. (1983). *Teaching the spoken language* (1st ed.). Cambridge University Press.

Brown, J. D. (1991). Do English and ESL faculties rate writing samples differently? *TESOL Quarterly*, *25*(4), 587–603. https://doi.org/10.2307/3587078

Brown, J. D. (1996). *Testing in language programs* (2nd ed.). Prentice Hall.

Brown, J. D., & Hudson, T. (2012). *Criterion-referenced language testing*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524803

Caban, H. L. (2003). Rater group bias in the speaking assessment of four L1 Japanese ESL students. *Second Language Studies, 21*(2), 1–44.

Cai, H. (2015). Weight-based classification of raters and rater cognition in an EFL speaking test. *Language Assessment Quarterly*, *12*(3), 262–282. https://doi.org/10.1080/15434303.2015.1053134

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing*, *28*(2), 201–219. https://doi.org/10.1177/0265532210393704

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing*, *12*(1), 16–33. https://doi.org/10.1177/026553229501200102

Chalhoub-Deville, M., & Wigglesworth, G. (2005). Rater judgment and English language speaking proficiency. *World Englishes*, *24*(3), 383–391. https://doi.org/10.1111/j.0083-2919.2005.00419.x

Chuang, Y. (2009). Foreign language speaking assessment: Taiwanese college English teachers' scoring performance in the holistic and analytic rating methods. *Asian EFL Journal*, *11*(1), 150-173.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). SAGE Publications, Inc.

Creswell, J. W. (2015). *A concise introduction to mixed methods research*. SAGE Publications, Inc.

Creswell, J. W., & Clark, V. P. L. (2017). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications, Inc.

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

Cumming, A., Kantor, R., & Powers, D. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype tasks: An investigation into raters' decision making and development of a preliminary analytic framework* (TOEFL Monograph Series, Report No: 22). Educational Testing Service. https://www.ets.org/research/policy_research_reports/publications/report/2001/icbg

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

Çoban, M. (2019, June 19). *A sample verbal protocol* [Video file]. Retrieved from https://youtu.be/IvN8rhDi8Fk

Daly, J. A., & Dickson-Markman, F. (1982). Contrast effects in evaluating essays. *Journal of Educational Measurement*, *19*(4), 309-316.

Davies, A. (2010). Test fairness: A response. *Language Testing*, *27*(2), 171–176. https://doi.org/10.1177/0265532209349466

Davis, L. (2012). *Rater expertise in a second language speaking assessment: The influence of training and experience* [Doctoral dissertation, University of Hawai'i at Manoa]. https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/cf030c00-6615-4d76-ad80-e5db3a33b9fd/content

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

Delaruelle, S. (1997). Text type and rater decision–making in the writing module. In G. Brindley & G. Wigglesworth (Eds.), *Access: Issues in language test design and delivery* (pp. 215–242). Sidney, Australia: National Centre for English Language Teaching and Research, Macquarie University.

Doane, D. P., & Seward, L. E. (2011). Measuring skewness: A forgotten statistic? *Journal of Statistics Education*, *19*(2). https://doi.org/10.1080/10691898.2011.11889611

Douglas, D. (1997). *Testing speaking ability in academic contexts: theoretical considerations*. (TOEFL Monograph Series No. 8). Educational Testing Service. https://www.ets.org/research/policy_research_reports/publications/report/1997/icia

Douglas, D. (2010). *Understanding language testing* (1st ed.). Routledge.

Douglas, D., & Myers, R. (2000). Assessing the communication skills of veterinary students: Whose criteria? In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida.* Cambridge University Press.

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative and mixed methodologies* (1st ed.). Oxford University Press.

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing*, *26*(3), 423–443. https://doi.org/10.1177/0265532209104669

Duff, P. (2008). *Case study research in applied linguistics* (1st ed.). Lawrence Erlbaum Associates.

Duijm, K., Schoonen, R., & Hulstijn, J. H. (2017). Professional and non-professional raters' responsiveness to fluency and accuracy in L2 speech: An experimental approach. *Language Testing*, *35*(4), 501–527. https://doi.org/10.1177/0265532217712553

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer? *Language Testing*, *19*(4), 347–368. https://doi.org/10.1191/0265532202lt235oa

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, *4*(2), 139–155. https://doi.org/10.1016/1060-3743(95)90004-7

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39. https://doi.org/10.2307/3151312

Fox, J. D. (2003). From products to process: An ecological approach to bias detection. *International Journal of Testing*, *3*(1), 21–47. https://doi.org/10.1207/s15327574ijt0301_2

Frederiksen, J. R. (1992). *Learning to "see:" Scoring video portfolios or "beyond the hunter-gatherer" in performance assessment*. [Paper presentation]. The Annual Meeting of the American Educational Research Association, San Francisco, CA.

Freedman, S. W. (1981). Influences on evaluators of expository essays: Beyond the text. *Research in the Teaching of English, 15*(3), 245-255. https://www.jstor.org/stable/40170792

Freedman, S. W., & Calfee, R. C. (1983). Holistic assessment of writing: Experimental design and cognitive and cognitive theory. In P. Mosenthal, L. Tamor, & S. A. Walmsley (Eds.), *Research on writing: Principles and methods* (pp. 75-98). Longman.

Fulcher, G. (2010). *Practical language testing* (1st ed.). Routledge.

Fulcher, G. (2014). *Testing second language speaking* (1st ed.) [E-book]. Routledge.

Fulcher, G. (2015). Assessing second language speaking. *Language Teaching*, *48*(2), 198–216. https://doi.org/10.1017/s0261444814000391

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book* (1st ed.). Routledge.

Fulcher, G., Davidson, F., & Kemp, J. (2010). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, *28*(1), 5–29. https://doi.org/10.1177/0265532209359514

Fulcher, G., & Reiter, R. M. (2003). Task difficulty in speaking tests. *Language Testing*, *20*(3), 321–344. https://doi.org/10.1191/0265532203lt259oa

Galaczi, E., & Taylor, L. (2018). Interactional competence: Conceptualizations, operationalizations, and outstanding questions. *Language Assessment Quarterly*, *15*(3), 219–236. https://doi.org/10.1080/15434303.2018.1453816

Gao, X., & Brennan, R. L. (2001). Variability of estimated variance components and related statistics in a performance assessment. *Applied Measurement in Education*, *14*(2), 191–203. https://doi.org/10.1207/s15324818ame1402_5

Gebril, A., & Plakans, L. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. *Assessing Writing, 21*(2), 56–73. https://doi.org/10.1016/j.asw.2014.03.002

Ginther, A. (2013). Assessment of speaking. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics.* Blackwell. http://dx.doi.org/%2010.1002/9781405198431

Goh, C. C. M., & Burns, A. (2012). *Teaching speaking: A holistic approach* (1st ed.). Cambridge University Press.

Green, A. (2014). *Exploring language assessment and testing* (1st ed.). Routledge.

Guetterman, T. C., & Mitchell, N. (2015). The role of leadership and culture in creating meaningful assessment: A mixed methods case study. *Innovative Higher Education*, *41*(1), 43–57. https://doi.org/10.1007/s10755-015-9330-y

Gui, M. (2012). Exploring differences between Chinese and American EFL teachers' evaluations of speech performance. *Language Assessment Quarterly*, *9*(2), 186-203. https://doi.org/10.1080/15434303.2011.614030

Han, Q. (2016). Rater cognition in L2 speaking assessment: A review of the literature. *Working Papers in TESOL & Applied Linguistics, 16*(1), 1-24.

Han, T. (2017). Scores assigned by inexpert raters to different quality of EFL compositions, and the raters' decision-making behaviors. *International Journal of Progressive Education, 13*(1), 136-152.

Henning, G. (1987). *A guide to language testing: Development evaluation research*. Longman ELT.

Hirai, A., & Koizumi, R. (2009). Development of a practical speaking test with a positive impact on learning using a story retelling technique. *Language assessment quarterly*, *6*(2), 151-167. https://doi.org/10.1080/15434300902801925

Hsieh, C. N. (2011). Rater effects in ITA testing: ESL teachers' versus American undergraduates' judgments of accentedness, comprehensibility, and oral proficiency. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *9*, 47-74.

Hsu, L. (2012). Applications of generalizability theory to estimate the reliability of EFL learners' performance-based assessment: A preliminary study. *Educational Research, 3*(2), 145-154.

Huang, B., Alegre, A., & Eisenberg, A. (2016). A Cross-Linguistic investigation of the effect of raters' accent familiarity on speaking assessment. *Language Assessment Quarterly*, *13*(1), 25–41. https://doi.org/10.1080/15434303.2015.1134540

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. https://doi.org/10.1016/j.system.2013.07.009

Huang, B. H., & Jun, S. A. (2015). Age matters, and so may raters: Rater differences in the assessment of foreign accents. *Studies in Second Language Acquisition*, *37*(4), 623–650. https://doi.org/10.1017/s0272263114000576

Huang, H. T. D., Hung, S. T. A., & Plakans, L. (2018). Topical knowledge in L2 speaking assessment: Comparing independent and integrated speaking test tasks. *Language Testing*, *35*(1), 27–49. https://doi.org/10.1177/0265532216677106

Hubbard, C., Gilbert, S., & Pidcock, J. (2006) Assessment processes in speaking tests: A pilot verbal protocol study, *Cambridge ESOL Research Notes*, *24*, 14-19.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge University Press.

Hughes, R. (2010). *Teaching and Researching Speaking* (2nd ed.). Longman. https://doi.org/10.4324/9781315833736

Isaacs, T. (2016). Assessing speaking. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 131–146). Berlin: DeGruyter Mouton. https://doi.org/10.1515/9781614513827-011

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

Iwashita, N. (1998). The validity of the paired interview in oral performance assessment. *Melbourne Papers in Language Testing*, *5*(2): 51-65.

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. https://doi.org/10.1111/0023-8333.00160

Kane, M. (2010). Validity and fairness. *Language Testing*, *27*(2), 177–182. https://doi.org/10.1177/0265532209349467

Kane, M. (2012). Articulating a validity argument. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 34–47). Routledge. 10.4324/9780203181287.ch2

Kang, O. (2008). *Ratings of L2 oral performance in English: relative impact of rater characteristics and acoustic measures of accentedness* [Doctoral dissertation, University of Georgia]. http://getd.libs.uga.edu/pdfs/kang_okim_200805_phd.pdf

Kang, O., Rubin, D., & Kermad, A. (2019). The effect of training and rater differences on oral proficiency assessment. *Language Testing*, *36*(4), 481–504. https://doi.org/10.1177/0265532219849522

Khabbazbashi, N. (2016). Topic and background knowledge effects on performance in speaking assessment. *Language Testing*, *34*(1), 23–48. https://doi.org/10.1177/0265532215595666

Kim, H. J. (2006). Issues of rating scales in speaking performance assessment. *Studies in Applied Linguistics and TESOL*, *6*(2). https://doi.org/10.7916/salt.v6i2.1549

Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. *Language Assessment Quarterly*, *12*(3), 239-261. https://doi.org/10.1080/15434303.2015.1049353

Kim, J., & Craig, D. A. (2012). Validation of a video conferenced speaking test. *Computer Assisted Language Learning*, *25*(3), 257–275. https://doi.org/10.1080/09588221.2011.649482

Kim, Y. H. (2009a). A G-Theory analysis of rater effect in ESL speaking assessment. *Applied Linguistics*, *30*(3), 435–440. https://doi.org/10.1093/applin/amp035

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

Kitchen, H., Bethell, G., Fordham, E., Henderson, K., & Ruochen, L. R. (2019). *OECD Reviews of Evaluation and Assessment in Education: Student Assessment in Turkey, OECD Reviews of Evaluation and Assessment in Education*, OECD Publishing. https://doi.org/10.1787/5edc0abe-en

Knight, B. (1992). Assessing speaking skills: a workshop for teacher development. *ELT Journal, 46*(3), 294-302. https://doi.org/10.1093/elt/46.3.294

Kopriva, R. (2008). *Improving testing for English language learners* (1st ed.). Routledge.

Kuiken, F., & Vedder, I. (2014). Raters' decisions, rating procedures and rating scales. *Language Testing*, *31*(3), 279–284. https://doi.org/10.1177/0265532214526179

Kyngäs H. (2020). Inductive Content Analysis. In Kyngäs H., Mikkonen K., Kääriäinen M. (Eds.) *The Application of Content Analysis in Nursing Science Research* (pp.13-21). Springer. https://doi.org/10.1007/978-3-030-30199-6_2

Lam, L. W. (2012). Impact of competitiveness on salespeople's commitment and performance. *Journal of Business Research*, *65*(9), 1328–1334. https://doi.org/10.1016/j.jbusres.2011.10.026

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Leckie, G., & Baird, J. A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement, 48*(4), 399-418. https://www.jstor.org/stable/41427532

Lee, Y. W. (2005). *Dependability of scores for a new ESL speaking test: Evaluating prototype tasks.* (TOEFL Monograph Series MS-28). Educational Testing Service. https://origin-www.ets.org/Media/Research/pdf/RM-04-07.pdf

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

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. https://doi.org/10.1177/0265532214530699

Longman. (n.d.). Experienced. In *ldoceonline.com dictionary*. Retrieved August, 2021, from https://www.ldoceonline.com/dictionary/experienced

Lumley, T. (1998). Perceptions of language-trained raters and occupational experts in a test of occupational English language proficiency. *English for Specific Purposes*, *17*(4), 347–367. https://doi.org/10.1016/s0889-4906(97)00016-1

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

Lumley, T., & McNamara, T. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*(1), 54–71. https://doi.org/10.1177/026553229501200104

Lumley, T., & O'Sullivan, B. (2005). The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking. *Language Testing*, *22*(4), 415–437. https://doi.org/10.1191/0265532205lt303oa

Luoma, S. (2004). *Assessing Speaking*. Cambridge University Press.

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design.* Lawrence Erlbaum Associates Publishers.

McNamara, T. F. (1996). *Measuring second language performance* (1st ed.). Longman Pub Group.

McNamara, T. F. (1997). "Interaction" in second language performance assessment: Whose performance? *Applied Linguistics*, *18*(4), 446–466. https://doi.org/10.1093/applin/18.4.446

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

Madsen, H. S. (1983). *Techniques in Testing: Teaching Techniques in English as a Second Language* (1st ed.). Oxford University Press.

Maxwell, J. A., & Loomis, D. (2003). Mixed method design: An alternative approach. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 241–271). Sage Publication Inc.

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing*, *26*(3), 397-421. https://doi.org/10.1177/0265532209104668

May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly*, *8*(2), 127-145. https://doi.org/10.1080/15434303.2011.565845

Myford, C. M., Marr, D. B., & Linacre, J. M. (1996). *Reader Calibration and Its Potential Role in Equating for the Test of Written English.* (TOEFL Research Report No. 52, RR-95-40). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1995.tb01674.x

North, B. (2012). *The development of a common framework scale of language proficiency*. Peter Lang Inc., International Academic Publishers.

Norton, J. (2005). The paired format in the Cambridge speaking tests. *ELT journal*, *59*(4), 287-297. https://doi.org/10.1093/elt/cci057

Nunan, D. (1989). *Designing tasks for the communicative classroom*. Cambridge University Press.

Ockey, G. J., Timpe-Laughlin, V., Davis, L., & Gu, L. (2019). Exploring the potential of a Video-Mediated interactive speaking assessment. *ETS Research Report Series*, *2019*(1), 1–29. https://doi.org/10.1002/ets2.12240

Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret test scores. *System*, *30*(2), 143–154. https://doi.org/10.1016/s0346-251x(02)00002-7

O'Sullivan, B. (2013). Assessing speaking. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 156–71). John Wiley & Sons. https://doi.org/10.1002/9781118411360.wbcla084

Papajohn, D. (2002). Concept mapping for rater training. *TESOL Quarterly*, *36*(2), 219–233. https://doi.org/10.2307/3588333

Pollitt, A., & Murray, N. L. (1996). What raters really pay attention to. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium (LTRC), Cambridge and Arnhem* (pp. 74- 91). Cambridge University Press.

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of statistical modeling and analytics*, *2*(1), 21-33.

Richards, J.C., & Schmidt, R.W. (2013). *Longman Dictionary of Language Teaching and Applied Linguistics* (4th ed.). Routledge. https://doi.org/10.4324/9781315833835

Ross, S. J. (2012). Claims, evidence, and inference in performance assessment. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 223-233). Routledge. https://doi.org/10.4324/9780203181287

Sakyi, A. A. (2003). *A study of the holistic scoring behaviors of experienced and novice ESL instructors*. [Doctoral dissertation, University of Toronto]. Library and Archives Canada. https://central.bac-lac.gc.ca/.item?id=NQ78033&op=pdf&app=Library&oclc_number=55106099

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory* (1st ed.). SAGE Publications.

Shohamy, E. (1995). Performance assessment in language testing. *Annual Review of Applied Linguistics*, *15*, 188–211. https://doi.org/10.1017/s0267190500002683

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.

Song, B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking, and ESL students? *Journal of Second Language Writing, 5*(2), 163-182. https://doi.org/10.1016/S1060-3743(96)90023-5

Stitt, J. K., Simonds, C. J., & Hunt, S. K. (2003). Evaluation fidelity: An examination of criterion-based assessment and rater training in the speech communication classroom. *Communication Studies*, *54*(3), 341–353. https://doi.org/10.1080/10510970309363290

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Second edition: Techniques and procedures for developing grounded theory* (2nd ed.). SAGE Publications, Inc.

Suen, H. K. (1990). *Principles of test theories* (1st ed.). Routledge.

Şahan, Ö. (2018). *The impact of rater experience and essay quality on rater behavior and scoring*. [Doctoral dissertation, Çanakkale Onsekiz Mart University]. Council of Higher Education Thesis Center https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Şahan, Ö. (2019). The impact of rater experience and essay quality on the variability of EFL writing scores. In S. Papageorgiou & K. M. Bailey (Eds.), *Global perspectives on language assessment: Research, theory, and practice* (pp. 32–46). Routledge. https://doi.org/10.4324/9780429437922

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

Swain, M. (2001). Examining dialogue: another approach to content specification and to validating inferences drawn from test scores. *Language Testing, 18*(3), 275-302. https://doi.org/10.1177/026553220101800302

Tashakkori, A., & Teddlie, C. (2003). The past and future of mixed methods research: From data triangulation to mixed model designs. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social & behavioral research* (pp. 671-702). Sage Publishing.

Taşdelen Teker, G., & Güler, N. (2019). Thematic content analysis of studies using generalizability theory. *International Journal of Assessment Tools in Education*, *6*(2), 279–299. https://doi.org/10.21449/ijate.569996

Tavakoli, P. (2009). Investigating task difficulty: learners' and teachers' perceptions. *International Journal of Applied Linguistics*, *19*(1), 1-25. https://doi.org/10.1111/j.1473-4192.2009.00216.x

Teng, H-C. (2007). A study of task type for L2 speaking assessment. [Paper presentation]. The Annual Meeting of the International Society for Language Studies (ISLS), Honolulu, HI. https://files.eric.ed.gov/fulltext/ED496075.pdf

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal*, *49*(1), 3–12. https://doi.org/10.1093/elt/49.1.3

Van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly*, *23*(3), 489–508. https://doi.org/10.2307/3586922

Vaughan, C. (1991). Holistic assessment: What goes on in the raters' minds? In. L. HampLyons (Ed.), *Assessing second language writing in academic contexts* (pp. 11-126). Van Haren Publishing.

Wang, B. (2010). On rater agreement and rater training. *English Language Teaching*, *3*(1), 108–112. https://doi.org/10.5539/elt.v3n1p108

Wei, J., & Llosa, L. (2015). Investigating differences between American and Indian raters in assessing TOEFL iBT speaking tasks. *Language Assessment Quarterly*, *12*(3), 283–304. https://doi.org/10.1080/15434303.2015.1037446

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

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

Weir, C. (2005). *Language testing and validation*. Palgrave Macmillan.

Weir, C. J., & Wu, J. R. (2006). Establishing test form and individual task comparability: a case study of a semi-direct speaking test. *Language Testing*, *23*(2), 167–197. https://doi.org/10.1191/0265532206lt326oa

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, *10*(3), 305–319. https://doi.org/10.1177/026553229301000306

Winke, P. (2012). Rating oral language. In C. A. Chapelle (Ed.), *Encyclopedia of applied linguistics* (pp. 4849–4855). Wiley. https://doi.org/10.1002/9781405198431.wbeal0993.pub2

Winke, P., Gass, S., & Myford, C. (2012). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*(2), 231–252. https://doi.org/10.1177/0265532212456968

Wolfe, E. F., & Feltovich, B. (1994). *Learning to rate essays: A study of scorer cognition.* [Paper presentation]. The Annual Meeting of the American Educational Research Association, New Orleans, LA. https://files.eric.ed.gov/fulltext/ED368777.pdf

Wolfe, E. W. (2005). Uncovering rater's cognitive processing and focus using think-aloud *protocols. Journal of Writing Assessment, 2*(1), 37-56. https://escholarship.org/uc/item/83b618ww

Wolfe, E. W., Kao, C., & Ranney, M. (1998). Cognitive differences in proficient and non-proficient essay scorers. *Written Communication*, *15*(4), 465-492. https://doi.org/10.1177/0741088398015004002

Xi, X. (2007). Evaluating analytic scoring for the TOEFL® Academic Speaking Test (TAST) for operational use. *Language Testing*, *24*(2), 251-286. https://doi.org/10.1177/0265532207076365

Xi, X., & Mollaun, P. (2009). How do raters from India perform in scoring the TOEFL iBT™ speaking section and what kind of training helps? *ETS Research Report Series*, *2009*(2), 1–37. https://doi.org/10.1002/j.2333-8504.2009.tb02188.x

Xi, X., & Mollaun, P. (2011). Using raters from India to score a Large-Scale speaking test. *Language Learning*, *61*(4), 1222–1255. https://doi.org/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. https://doi.org/10.1177/0265532214536171

Yin, R. K. (2014). *Case study research: Design and methods* (5th ed.). SAGE Publications, Inc. https://doi.org/10.1177/0265532209360671

APPENDICES

APPENDIX A

RATER PROFILE FORM

The purpose of this questionnaire is to collect background information for my dissertation study entitled *"The Effect of Rater Experience and L2 Speaking Performance Quality on Score Variation and Rater Behavior"*. Your information and identity will be kept confidential. Your participation in this study is entirely voluntary. You have the right not to fill in the questionnaire. Leaving the study will not result in any penalty or affect your relations with your institution.

I would like to thank you for your cooperation and contribution to this study.

1. **Your Name (Pseudonym) is…….**

2. **Your gender:**

Male……. Female…….

3. **Your age:**

……

4. **What is your highest level of education?**

B.A……... M.A…… Ph.D……... Other, please specify……...

5. **I have been teaching *EFL* (English as a Foreign Language) for …… years.**

6. **I have been teaching *EFL* at the *university* for …… years.**

7. **Have you ever taught any *EFL speaking/communication courses* at the *university?***

Yes (…) No (…)

If *yes,* please fill out the necessary information below (The number of speaking/communication classes in a quarter term and academic year in *your current institution*)

(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes

If *yes,* please fill out the necessary information below (The number of speaking/communication classes in a quarter term and academic year at *other higher education institutions*)

(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes

(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes
(……) academic year ……quarter term (……)the number of speaking classes

8. The *number* of speaking assessment duties in *your current institution* *(only test sessions) (not performance tasks or assignments)*

……. times

9. The *number* of speaking assessment duties at *other higher education institutions* *(only tests) (not performance tasks or assignments)*

……. times

10. The *number* of *trainings* in speaking assessment *(a formal professional development session based on speaking assessment)*

……. times

11. How would you *describe* your experience as an *EFL speaking rater*?

I have no experience …...…

I have little experience …...…

I have some experience …...…

I am experienced …...…

I am very experienced …...…

APPENDIX B

ANALYTIC SCORING RUBRIC

CATEGORY	EXEMPLARY-4	ACCOMPLISHED-3	LIMITED-2	UNSATISFACTORY-1	SCORE
1. TASK COMPLETION/ CONTENT	Thorough & lengthy answers. Relevant & appropriate answers. Organized response with plenty of details & examples.	Adequate answers in length and task completion. Mostly relevant answers. Some organization, details & examples.	Mostly short answers with almost no elaboration. Some wandering off the topic but some relevant answers. Insufficient organization, details & examples.	Short answers with no effort for elaboration. Irrelevant answers. No organization, details & examples.	…… x 8 = ……
2. VOCABULARY	Wide range of vocabulary & expressions. Appropriate word choice. Attempts at going beyond expected vocabulary.	Sufficient vocabulary & expressions. Good word choice with few mistakes. Expected vocabulary level demonstrated.	Limited variation in vocabulary and expressions. Occasional mistakes in word choice. Insufficient vocabulary level impeding communication.	Basic, below expected vocabulary & expressions. Major mistakes in word choice resulting in miscommunication.	…… x 5 = ……

3. GRAMMAR & STRUCTURE	Sound & varied grammatical structures. Almost no grammar mistakes. Word order & sentence structure in place.	Good grammar with less variation. Few minor grammar mistakes. Mostly correct word order & sentence structure.	Limited grammatical structures. Some grammar mistakes resulting in incompetent conversation. Major recurring mistakes in word order & sentence structure.	Insufficient grammar to convey ideas. Lack of word order & sentence structure awareness.	…… x 6 = ……
4. FLUENCY	Almost no hesitation, fillers or searching for words. Student in control, smooth & at ease. Almost no repetition, self-correction or fragmentary language. Accurate & clear pronunciation.	Little hesitation, rare usage of fillers or searching for words. Some lagging. Few repetitions, self-corrections or little fragmentary language. Intelligible pronunciation with minor mistakes.	Continuous and recurring hesitation and fillers. Frequent lagging. A lot of repetitions, self-corrections or searching for words. Unintelligible pronunciation often hindering meaning.	No sign or effort for fluency observed. Constant lagging. Unclear pronunciation.	…… x 6 = ……
					Total .../100

APPENDIX C

RATER EXPERIENCE SCALE FORM

Rater Name:	
Assessment experience (60%)	
Assessing speaking duties in your current institution (……) Assessing speaking duties at other higher education institutions (……)	
Teaching experience (30%)	
Teaching speaking experience in your current institution***** (……) Teaching speaking experience at other higher education institutions ***** (……)	
Training experience (10%)	
Speaking assessment training sessions (……)	
Total	

Each **speaking/communication class in one quarter term equals to **1 point.***

APPENDIX D

ASSESSMENT INSTRUCTIONS FOR QUALITY CHECK RATERS

Dear Quality Rater,

I am a PhD candidate at Çanakkale Onsekiz Mart University in Department of ELT. The purpose of this dissertation is to explore the effect of rater experience and L2 speaking performance quality on score variation and rater behavior. In this regard, you are kindly requested to evaluate the L2 speaking performances and categorize them based on their relevant quality. This division is crucial for the main data collection of this research. Please pay attention to the following items while evaluating the L2 speaking performances provided to you. I would like to express my sincere gratitude to all of you for your kind contribution.

Mustafa ÇOBAN
PhD Candidate
Çanakkale Onsekiz Mart University
e-mail:

- Please read the speaking exam tasks before assessing the L2 speaking performances. Make sure you keep them on your desk while grading.

- The L2 speaking performances were collected from the official speaking exam produced by preparatory program university EFL students with B1 language proficiency level.

- You are expected to score the L2 speaking performances. Use the analytic rubric to determine the quality of the speaking performance as low-quality, medium-quality, and high-quality.

- Please evaluate the L2 speaking performances individually rather than comparing their quality to each other.

- Please use only the analytic rubric for each L2 speaking performance rather than relying on personal opinions.

APPENDIX E

INSTRUCTIONS FOR ASSESSMENT AND VERBAL PROTOCOLS

Please examine these instructions carefully before you start to rate the L2 speaking performances.

Purpose

These instructions are written to help guide you and others in producing think-aloud protocols for this project in a consistent and informative manner. Retrospective think-aloud protocols ask people to say everything they thought while they performed a task in order to document and better understand what raters paid attention to and considered important when they completed a task. The purpose of the think-aloud protocols for this study is to find out in as much detail as possible what you can actually remember thinking, deciding, and doing while rating a sample of L2 speaking performances. The most important thing to remember is to say everything you thought, and to make certain this is recorded clearly onto the voice-recorder. What you say will become important data for my dissertation. Thanks in advance.

The Assessment Task

You will receive a package of 60 L2 speaking performances produced by preparatory program university EFL students with B1 language proficiency level. While you will assess 15 of them using think-aloud protocol, you will follow standard procedures for the other 45 L2 speaking performances. The L2 speaking performances that you will utilize thinking aloud are shown in the recording files called 'USE THINK-ALOUD PROTOCOL SPXXX'. You will also find the list of those L2 speaking performances in this document. Say as much as you can what you remember while and after you listen to the speaking performance and decide on how to rate it.

The L2 speaking performances

You will also receive copies of the speaking performance prompt originally given to the students so you know what they were asked to give response. There are only three tasks and ten questions in each task. The L2 speaking performances have been identified with code numbers. The order in which you receive the L2 speaking performances has been sequenced randomly in quality, but you should receive L2 speaking performances varying qualities.

The Ratings

In making your assessment, try to use 'Analytic Speaking Exam Rubric' as the basis for your decision. The rating will not be judged as right or wrong. However, I will be analyzing the scores that you assign to the L2 speaking performances along with the spoken data regarding your thoughts while assessing the L2 speaking performances.

Recording Your Thoughts While Assessing

- Keep talking, conveying your thoughts continuously while you assess the L2 speaking performances beginning from the moment you first hear the speaking performance until you have completed rating it.
- Feel free to speak in either **English** or **Turkish.** If you speak in Turkish, it will be translated into English for the final analysis of the data.
- Speak continuously. Report fully, even what might seem trivial. Do not assume that others know what you are doing or thinking.
- Try to avoid speech fillers (i.e., uh, um) as much as possible. Try to use words instead, so that I can understand what your thoughts are.
- Talk and make your assessment as naturally and as honestly as you can, according to what you usually do when you assess students' L2 speaking performances. Don't start rationalizing your ideas at length; I am just interested in your natural thought process as you made decisions.

Instructions for Recording

1. Turn on the voice-recorder or your smart phone so that can record your voice and check that it works. Check whether it records properly and that the quality of the recording is clear by trying out a few words initially, then playing it back. Make sure there is no background noise (e.g., fans, music, foot tapping, etc.).
2. Keep the recorder/smart phone at an appropriate distance from your face and be sure it captures your voice clearly.
3. Turn on the recorder/smart phone, and state the speaking performance code and your name (or pseudonym to be used in the research) at the beginning of each speaking performance.

4. While rating the speaking performance, follow the instructions above **(*Recording Your Thoughts While Assessing).*** Then, when you have made a rating decision, indicate the score that you have assigned to the speaking performance.
5. You will write three reasons that impact you most for your decision about the speaking performance. Feel free to write other notes on the rubric if you like, but I will not be analyzing your written notes.
6. If you happen to reconsider any of your ratings (e.g., for a second or third time), verbalize your reason(s) for doing so and indicate on the recorder that this is what you are doing.
7. If you have to take a break while you are assessing the L2 speaking performances, indicate on the recorder that you are doing this, turn the voice recorder off or pause the recording on your smart phone. Then, when you start again, indicate this clearly on the device.
8. When you have completed assessing the speaking performance, turn off the voice-recorder/smart phone.
9. Please record your thoughts for each speaking performance separately.
10. At the end of the assessment session and voice-recording, please put all the L2 speaking performances together with the recorder that you used (if you were provided with one) back into the package. Thank you!

(adapted from Cumming, Kantor & Powers, 2001, pp. 83-85)

THINK-ALOUD PROTOCOL WILL BE USED FOR THE FOLLOWING L2 SPEAKING PERFORMANCES:

SP001	SP020	SP038
SP003	S0021	SP044
SP007	SP023	SP052
SP012	SP029	SP053
SP019	SP034	SP060

APPENDIX F

CODING SCHEME FOR DECISION-MAKING BEHAVIORS

A. SELF-MONITORING FOCUS
Self-Monitoring Focus-Interpretation Strategies 1. Interpret speaking performance prompt or test items 2. Consider personal situation of the test takers 3. Refer to scoring rubric
Self-monitoring Focus-Judgment Strategies 1. Evaluate responses in comparison with other benchmarks or responses 2. State overall performance of the test takers *3.* State or revisit scoring
B. RHETORICAL AND IDEATIONAL FOCUS
Rhetorical and Ideational Focus-Interpretation Strategies 1. Interpret vague or equivocal expressions 2. Restate test takers' ideas or propositions
Rhetorical and Ideational Focus-Judgment Strategies 1. Evaluate topic development 2. Evaluate task completion, content and relevance 3. Evaluate originality and creativity 4. Recognize unnecessary or verbose expressions 5. Evaluate organization of the response *6.* Evaluate register of the test takers

C. LANGUAGE FOCUS

Language Focus-Interpretation Strategies

1. Group errors into types
2. Rephrase responses for interpretation

Language Focus-Judgment Strategies

1. Evaluate intelligibility of the response
2. Consider errors in terms of quantity and frequency
3. Evaluate fluency
4. Evaluate vocabulary
5. Rate overall language use
6. Evaluate accent or pronunciation
7. Evaluate grammar and sentence structures
8. Evaluate L1 use of the test takers

(adapted from Cumming, Kantor, & Powers, 2002, pp. 93-94)

APPENDIX G

DESCRIPTIVE STATISTICS FOR SCORES ASSIGNED TO HIGH-QUALITY L2 SPEAKING PERFORMANCES

Performance	Range	Minimum	Maximum	*M*	*SD*
1	19	81	100	93.83	5.08
2	26	63	89	76.71	7.18
3	28	73	100	93.91	6.25
4	32	63	95	81.68	9.28
5	32	63	95	83.68	8.01
6	29	66	94	78.68	8.10
7	36	61	97	78.88	8.48
8	29	69	98	84.85	8.14
9	34	67	100	85.08	8.76
10	22	70	92	81.44	5.60
11	36	61	97	77.72	9.70
12	29	60	89	76.96	6.91
13	28	73	100	86.10	7.57
14	26	71	97	83.45	8.08
15	28	70	97	81.80	7.66
16	24	77	100	87.63	6.50
17	33	68	100	85.47	8.23
18	25	70	94	82.22	6.81
19	28	72	100	85.26	7.37
20	25	72	97	86.31	7.34

APPENDIX H

DESCRIPTIVE STATISTICS FOR SCORES ASSIGNED TO MEDIUM-QUALITY L2 SPEAKING PERFORMANCES

Performance	Range	Minimum	Maximum	*M*	*SD*
21	34	61	95	76.49	8.59
22	25	75	100	86.24	7.50
23	32	60	92	74.58	8.84
24	31	54	85	68.59	7.92
25	27	71	98	82.98	7.19
26	29	57	86	74.47	7.72
27	27	63	89	76.15	7.43
28	35	60	95	80.11	9.13
29	25	61	86	72.64	6.86
30	36	56	92	71.79	8.52
31	34	50	84	70.04	6.64
32	42	47	89	70.90	10.6
33	20	63	83	72.13	6.38
34	32	50	82	67.26	7.85
35	40	46	86	68.75	9.23
36	31	60	90	77.65	9.11
37	20	70	89	77.65	5.33
38	27	61	87	72.06	6.54
39	24	62	86	75.73	6.23
40	36	52	88	66.40	8.31

APPENDIX I

DESCRIPTIVE STATISTICS FOR SCORES ASSIGNED TO LOW-QUALITY L2 SPEAKING PERFORMANCES

Performance	Range	Minimum	Maximum	*M*	*SD*
41	44	25	69	50.49	11.283
42	35	50	85	63.58	9.218
43	36	50	86	70.21	10.165
44	31	47	78	64.81	8.860
45	35	44	79	65.94	8.206
46	43	54	97	71.55	10.977
47	42	43	84	62.20	10.501
48	36	33	70	54.31	10.988
49	27	47	74	62.64	6.662
50	36	38	73	58.72	8.844
51	33	40	72	53.74	9.688
52	20	45	64	54.56	6.013
53	36	41	77	60.90	11.105
54	45	30	75	63.15	10.428
55	38	43	81	58.96	8.594
56	30	36	66	51.81	8.280
57	54	25	79	55.97	11.249
58	40	35	74	59.11	10.177
59	41	53	94	69.79	10.951
60	28	47	75	58.90	7.528

APPENDIX J
ETHICS COMMITTEE APPROVAL

T.C.
ÇANAKKALE ONSEKİZ MART ÜNİVERSİTESİ
SOSYAL BİLİMLER VE EĞİTİM BİLİMLERİ ETİK KURULU

PROJE/ARAŞTIRMA DEĞERLENDİRME SONUÇ RAPORU

Toplantı Tarihi	20. 09. 2018
Toplantı Sayısı	5
Başvuru protokol numarası	2018/49
Başvuru tarihi	18.09.2018
Proje/araştırma başlığı	Puanlayıcı Tecrübesi ve Sözlü Yanıt Kalitesinin Puan Değişkenliği ve Puanlayıcı Davranışı Üzerindeki Etkisi
Proje/araştırma yürütücüsü	Mustafa ÇOBAN
Karar	Bilimsel araştırma etik kurallarına uygundur.
Açıklamalar	----------