RATER EFFECTS AND FOCI
IN THE ORAL ASSESSMENT PROCESS OF ENGLISH TEST
FOR AVIATION PERSONNEL

ELİF RAHTUVAN

BOĞAZİÇİ UNIVERSITY
2021

RATER EFFECTS AND FOCI
IN THE ORAL ASSESSMENT PROCESS OF ENGLISH TEST
FOR AVIATION PERSONNEL

Thesis submitted to the

Institute for Graduate Studies in Social Sciences

in partial fulfillment of the requirements for the degree of

Master of Arts

in

English Language Education

by

Elif Rahtuvan

Boğaziçi University

2021

Rater Effects and Foci

in the Oral Assessment Process of English Test for Aviation Personnel

The thesis of Elif Rahtuvan

has been approved by:

Prof. Gülcan Erçetin _____
(Thesis Advisor)

Assoc. Prof. Sibel Tatar _____

Asst. Prof. Aylin Ünaldı _____
(External Member)

June 2021

DECLARATION OF ORIGINALITY

I, Elif Rahtuvan, certify that

- I am the sole author of this thesis and that I have fully acknowledged and documented in my thesis all sources of ideas and words, including digital resources, which have been produced or published by another person or institution;
- this thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;
- this is a true copy of the thesis approved by my advisor and thesis committee at Boğaziçi University, including final revisions required by them.

Signature.....

Date .....

ABSTRACT

Rater Effects and Foci

in the Oral Assessment Process of English Test for Aviation Personnel

Miscommunication in global aeronautical transmissions may lead to severe incidents and accidents in the aviation sector. In order to maintain flawless communication, International Civil Aviation Organization (ICAO) introduced Language Proficiency Requirements (LPR) and a Proficiency Rating Scale to aviation English test holders in 2003 (Alderson, 2009). However, several aviation English tests fail to provide detailed information about their reliability and validity (Alderson, 2011). Thus, this study intends to present evidence for rater effects and the areas they focus in the oral assessment part of English Test for Aviation Personnel (ETAP). Based on a theoretical model of a language for specific purposes (LSP) system by O'Sullivan and Weir (2011), the accuracy of raters' decisions in ETAP was investigated. Four ETAP raters participated in the study and evaluated eight pilots' oral performances while they were taking an eight-week-long online rater training. The assigned scores by ETAP raters were analyzed via Multi-Faceted Rasch Model (MFRM) (Linacre, 1989) by concentrating on rater severity/leniency, consistency, bias and central tendency effect. Through qualitative data including the raters' verbal reports, the study also aimed to explore ETAP raters' foci when evaluating oral performances. The results demonstrated raters' rating consistency with rare rater effects and high rater reliability. Besides, six common foci were drawn from the study that can be utilized in future rater training sessions in order to improve ETAP, its testing process and ETAP rating scale.

ÖZET

Havacılık Personeli için İngilizce Sınavının Sözlü Değerlendirme Sürecinde Değerlendirici Etkileri ve Odakları

Havacılık sektöründe muhataplar arasındaki yanlış iletişim, geri dönülmez sonuçlar doğuran ciddi kazalara neden olmaktadır. Uluslararası Sivil Havacılık Örgütü (ICAO), kazaları engellemek ve kusursuz iletişimi korumak için 2003 yılında Dil Yeterlilik Gereksinimleri (LPR) ve Yeterlilik Derecelendirme Ölçeğini yayınladı (Alderson, 2009). Havacılık İngilizcesi test sahipleri yayınlanan dokümanları göz önünde bulundurduklarını iddia etseler de birçok havacılık İngilizcesi testinde güvenirlik ve geçerlilik çalışmalarının eksikliği tespit edildi (Alderson, 2011). Bu durum göze alınarak, bu araştırmada Havacılık Personeli için İngilizce Sınavının (ETAP) sözlü değerlendirme sürecinde oluşan değerlendirici etkilerini ve odaklarını ortaya çıkarmak amaçlandı. O'Sullivan ve Weir (2011) tarafından geliştirilmiş olan özel amaçlar için dil (LSP) sisteminin teorik modeline dayanarak, değerlendiricilerin ETAP'taki kararlarının doğruluğu araştırıldı. Çalışmaya dört ETAP değerlendiricisi katıldı. Katılımcılar sekiz haftalık çevrimiçi değerlendirici eğitimi alırken sekiz pilotun sözlü performanslarını değerlendirdiler. Değerlendirici şiddetini ve hoşgörüsünü, tutarlılığını, yanlılığını ve merkezi eğilim etkilerini analiz etmek için çok yönlü Rasch modelinden (MFRM) (Linacre, 1989) yararlanıldı. Çalışmada ayrıca, ETAP puanlayıcılarının sözlü raporları da analiz edildi. Analiz sonuçları, nadir değerlendirici etkileri ve yüksek değerlendirici güvenilirliği ile puanlayıcıların değerlendirmedeki tutarlılıklarını gösterdi. Ek olarak, değerlendiricilerin ETAP sınavıyla ilgili altı ortak odak noktası tespit edildi.

ACKNOWLEDGEMENTS

First and foremost, I would like to present my gratitude to my thesis supervisor Prof. Gülcan Erçetin for spending her precious time to provide me with constructive feedback for my thesis. Furthermore, I owe many thanks to Asst. Prof. Aylin Ünaldı for sparing her valuable time to share her knowledge regarding assessment in aviation English with me. I also would like to express my gratitude to the committee member Assoc. Prof. Sibel Tatar for her feedback, kind words and guidance during my master's degree. Besides, I want to thank Prof. Ayşe Gürel for her guidance and help. She dealt with certain official documents for me to be able to continue my master's degree during my absence term in the school.

I would like to state my special gratefulness to Col. Sabit Çetin and Maj. Ayşe Terzi Altıparmak for their support. They shared their precious time with me by narrating their academic experiences and encouraged me to continue my academic career despite the heavy burden at work. I also present my thankfulness to Derya Önder and other sincere members of Turkish Foundation of Civil Aviation Pilots (PILVAK) for responding to my questions and for their help anytime I needed.

I would like to present my deepest gratitude to Asst. Prof. Vahid Aryadoust to make this thesis real. I benefited from his YouTube videos explaining how to conduct Rasch analysis in order to analyze my data for the study. I am very thankful to him for sharing his advanced statistics knowledge with everyone and answering my questions on YouTube.

Eventually, I want to state that I am very lucky to have such a great family who always supports and encourages me. I would like to thank them for their endless patience, unconditional love and support.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION 1
1.1 Introduction to the research 1
1.2 The scope of the research and its questions 3
1.3 Outline of the thesis 6
CHAPTER 2: LITERATURE REVIEW 7
2.1 Introduction 7
2.2 Assessment in ESP 7
2.3 Aviation English 13
2.4 ICAO 17
2.5 Rasch measurement 20
2.6 Rater effects 24
2.7 Rater training 27
2.8 Conclusion 30
CHAPTER 3: METHODOLOGY 32
3.1 Introduction 32
3.2 The purpose of the research 32
3.3 Participants 33
3.4 Data collection instruments 34
3.5 Procedures 40
3.6 Data analysis 41
3.7 Conclusion 46
CHAPTER 4: RESULTS 47
4.1 Research question 1: Do ETAP raters differ in their scoring in terms of their individual level of severity and leniency when assessing ETAP test takers' oral performances? 47
4.2 Research question 2: How consistently do ETAP raters rate ETAP test takers' oral performances? 49
4.3 Research question 3: Do ETAP raters show any central tendency effect due to individual indistinguishability of score levels in ETAP rating scale? 50
4.4 Research question 4: Do ETAP raters show any rater bias towards any criteria in the rating scale? If yes, for which criteria do the raters behave more severely or leniently? 59
4.5 Research question 5: What do ETAP raters focus while rating oral performances? 60

4.6 Conclusion 73
CHAPTER 5: DISCUSSION 74
5.1 Research question 1: Do ETAP raters differ in their scoring in terms of their individual level of severity and leniency when assessing ETAP test takers' oral performances? 74
5.2 Research question 2: How consistently do ETAP raters rate ETAP test takers' oral performances? 76
5.3 Research question 3: Do ETAP raters show any central tendency effect due to individual indistinguishability of score levels in ETAP rating scale? 77
5.4 Research question 4: Do ETAP raters show any rater bias towards any criteria in the rating scale? If yes, for which criteria do the raters behave more severely or leniently? 79
5.5 Research question 5: What do ETAP raters focus while rating oral performances? 82
5.6 Conclusion 87
CHAPTER 6: CONCLUSION 88
6.1 Introduction 88
6.2 Research implications 91
6.3 Limitations and suggestions for future research 92
APPENDIX A: ETAP RATING SCALES 94
REFERENCES 100

LIST OF TABLES

Table 1. Scoring Scheme for ETAP 38

Table 2. Descriptors for ETAP Raters' Verbal Reports 40

Table 3. Raters' Measurement Report 49

Table 4. Partial Credit Model Analysis for Criterion Structure in ETAP 52

Table 5. Partial Credit Model Analysis for Criterion Vocabulary in ETAP 53

Table 6. Partial Credit Model Analysis for Criterion Pronunciation in ETAP 54

Table 7. Partial Credit Model Analysis for Criterion Fluency in ETAP 54

Table 8. Partial Credit Model Analysis for Criterion Comprehension in ETAP 55

Table 9. Partial Credit Model Analysis for Criterion Interaction/Content in ETAP 56

Table 10. Deniz's Statistical Results for Criterion Structure 57

Table 11. Olcay's Statistical Results for Central Tendency Effect 58

Table 12. Test Takers' Measurement Report 59

Table 13. The First Bias/Interaction Pairwise Report 60

Table 14. The Second Bias/Interaction Pairwise Report 60

Table 15. Raters' Criterion Relevant and Irrelevant Comments 66

ABBREVIATIONS

ATC: Air Traffic Controllers

BUYEM: Boğaziçi University Lifelong Learning Centre

EAP: English for Academic Purposes

ELF: English as a Lingua Franca

ELPAC: English Language Proficiency for Aeronautical Communication

EMP: English for Medical Purposes

EOP: English for Occupational Purposes

EPTA: English Proficiency Test for Aviation

ESP: English for Specific Purposes

ETAP: English Test for Aviation Personnel

FAA: Federal Aviation Administration

FLIP: Flight Information Publication

ICAO: International Civil Aviation Organization

LPR: Language Proficiency Requirements

LSP: Language for Specific Purposes

MFRM: Multi-Faceted Rasch Model or Many Facet Rasch Measurement

MnSq: Mean Square

OET: Occupational English Test

PILVAK: Turkish Foundation of Civil Aviation Pilots

SE: Standard Error

STEP: Special Test of English Proficiency

TestDaF: Test of German as a Foreign Language

TLU: Target Language Use

UCLES: University of Cambridge Local Examinations Syndicate

ZStd: Z-standardized

CHAPTER 1

INTRODUCTION

1.1 Introduction to the research

English for specific purposes (ESP) is defined as a language teaching approach aiming to teach English to learners by basing teaching content and methods on learners' learning reasons and needs (Hutchinson & Waters, 1987). In the past, the need for teaching and learning ESP started with the commercial and technological developments requiring communication across languages (Paltridge & Starfield, 2013). With the international communication necessity in various professional domains, diverse areas of ESP such as English for occupational purposes (EOP), English for academic purposes (EAP) and English for medical purposes (EMP) have emerged.

In designing ESP courses, language needs of learners are prioritized. Therefore, the curriculum, course materials and assessment reflect particular domains in real life. In order to reflect real life domains, simulated circumstances that mirror real contexts are created for learners. This approach aims to guarantee an effective communication among interlocutors so that the intended service can be delivered safely and appropriately, i.e. in medical and aviation industries (Woodward-Kron & Elder, 2016). Therefore, assessing the English proficiency levels of candidates who plan to work in those areas has gained importance with the increasing popularity of ESP programs. Following the popularity of ESP assessment, authenticity in ESP tests was emphasized by Douglas (2001) with the focus on test format, language features and tasks that reflect real occupational context. With realization of problems regarding representativeness of ESP tests for real contexts, validation studies have

been conducted to detect the quality of different ESP tests (McNamara, 1990; Woodward-Kron & Elder, 2016).

Thanks to international mobility, aviation English has become an important ESP area for people working in the aviation sector such as pilots, air traffic controllers (ATC) and flight attendants. With the increasing necessity for standardized aviation English in the past decades, a great number of studies regarding the teaching of aviation English (Aiguo, 2008; Bullock, 2017; Cushing, 1994; Douglas, 2004, 2014; Kim, 2012; Kim & Elder, 2009, 2015) have been carried out. These studies documented the variability in application of aviation English language requirements across different aviation English tests, the integration of learning and testing process in aviation English, needs analysis for designing these programs, and significance of receiving expert opinions for aviation context.

Regarding assessment in aviation English, the International Civil Aviation Organization (ICAO) attempted to create standardization and published a document called Language Proficiency Requirements (LPR) and a Proficiency Rating Scale in 2003 (Alderson, 2009). According to ICAO, pilots and ATCs are required to have a certificate to document their proficiency in global aeronautical interaction in English. The language proficiency licenses provided to them are based on LPR which consists of six levels of proficiency. Test takers should obtain at least Level 4 for license and are subject to retesting in three years if their Level is 4. Besides, license holders who achieve Level 6 can have a lifetime license.

With the increasing need for language license in the aviation sector, many countries have developed their aviation tests. However, research has also documented the lack of evidence in validity and reliability of the aviation English tests (Alderson, 2009, 2010, 2011; Knoch, 2009). One of the major issues is

establishing the scorer reliability in performance-based speaking tests such as aviation English tests. In order to detect problems in rating reliability, bias and error analyses should be conducted. In this aspect, Popham (1990, as cited in Myford & Wolfe, 2003) states three different potential causes of errors in the rating process. These are (1) rating scales, (2) rating procedure, and (3) raters (Myford & Wolfe, 2003, p. 390). Regarding the raters, one of the common problems is that it can be challenging for raters to rate test performances in the same manner for each examinee throughout the rating process. Raters may show certain inconsistencies and bias while rating different performances for the same exam. These variances, such as severity/leniency, central tendency, inconsistency, halo effect and bias (Myford & Wolfe, 2003), in rating result in inconsistent and unfair scores being assigned to test takers. Since it is difficult to keep raters consistent in rating, despite rating trainings, the statistical analysis approach called many-facet Rasch measurement (MFRM) (Eckes, 2009) is a beneficial way to detect rater effects and to eliminate them by providing fair scores for test takers' performances.

1.2 The scope of the research and its questions

Aviation industry is one of the biggest industries across the world. Aviation related accidents can lead to unfortunate and irreversible consequences like loss of lives. Therefore, in order to prevent accidents resulting from miscommunication in global aeronautical transmissions, valid and reliable aviation English assessment process is considered vital. As for the standardization in aviation English assessment, ICAO sets certain language requirements. However, it neither develops aviation English tests nor monitors the current tests in the sector. Therefore, aviation English tests with different qualities have emerged. Firstly, one of the essential problems in

aviation English testing is whether tests are capable of assessing the relevant language constructs and skills in the aviation context. Secondly, the accuracy of decisions in performance-based speaking tests should be questioned.

In Turkey, Turkish Foundation of Civil Aviation Pilots (PILVAK) developed an aviation English test called English Test for Aviation Personnel (ETAP). Studies for ETAP task validation demonstrate that ETAP tasks represent construct relevant language usage in the target language domain (Ünaldı, 2019). However, evidence for scoring validity in ETAP still needs to be examined. Since scoring validity is significant for assigning fair scores to test takers, this study aims to investigate the accuracy of decisions in ETAP scoring process. Therefore, in this thesis, ETAP rating process is examined in terms of rater effects and raters' common foci when grading oral performances of test takers.

During forming a research idea for this thesis, studies about (1) language testing (Bachman & Palmer, 1996; Davies, 2001; Knoch, 2007; Lumley, 1998), (2) Rasch analysis for language test validation (Barkaoui, 2013; Du, Wright & Brown, 1996; Eckes, 2009; McNamara & Knoch, 2012; Myford & Wolfe, 2003, 2004), (3) rater effects and rating behaviors (Brown, 1995; Eckes, 2005; Kim, 2015; Knoch, Read, & Randow, 2007; Kuiken & Vedder, 2014; Lumley & McNamara, 1995) and (4) aviation English were analyzed. Then, a synthesis of those studies helped to create a research topic which aims to investigate rater effects and rating behaviors in an aviation English test through Rasch analysis. For the current study, eight test takers' scores assigned by four ETAP raters were utilized for Rasch measurement to find validity evidence for raters' scoring. Besides, raters' rating behaviors were examined via qualitative analysis of raters' verbal reports.

In the light of the brief information stated above, the following research questions were formed for the current study:

1. Do ETAP raters differ in their scoring in terms of their individual level of severity and leniency when assessing ETAP test takers' oral performances?
2. How consistently do ETAP raters rate ETAP test takers' oral performances?
3. Do ETAP raters show any central tendency effect due to individual indistinguishability of score levels in ETAP rating scale?
4. Do ETAP raters show any rater bias towards any criteria in the rating scale? If yes, for which criteria do the raters behave more severely or leniently?
5. What do ETAP raters focus while rating oral performances?

In order to answer the first four questions, Rasch analysis was conducted through the software program 'Minifac'. The first research question intends to find out whether raters are interchangeable or not due to their severity degrees in rating the oral performances. The second research question examines ETAP raters' reliability in terms of their rating consistency. In the third research question, whether raters overuse the middle categories in ETAP rating scale by avoiding the extreme categories is investigated. As for the fourth research question, rater bias towards any criteria in ETAP rating scale is analyzed. That is to say, whether raters rate a criterion more severely or leniently than the others is analyzed by detecting rater bias patterns in Rasch analysis tables. Lastly, the common focus points of raters when they rated examinees' oral performances are investigated by analyzing raters' verbal reports.

1.3 Outline of the thesis

After the first chapter including the introduction part of the study, Chapter 2 contains a review of related literature for the current research. The literature review covers main topics in ESP assessment, a theoretical model of a Language for Specific Purposes (LSP) system by O'sullivan and Weir (2011), aviation English including phraseology and plain language, ICAO, Rasch measurement, rater effects and rater training. In Chapter 3, the purpose of the study, research questions, participants, data collection instruments, procedure and data analysis are shared. Chapter 4 reports both quantitative and qualitative results in depth. Chapter 5 presents a thorough discussion for research findings. Finally, a conclusion summarizing the study, research implications, limitations and suggestions for future research are presented in Chapter 6.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This study investigates rater effects and foci in an oral assessment part of an aviation English test. Since the research topic is mainly based on assessment in aviation English, this chapter summarizes main points in ESP assessment including a theoretical model of an LSP system by O'sullivan and Weir (2011) to develop valid tests in LSP. Furthermore, aviation English as a part of ESP is introduced and two specific parts of aviation English language are explained briefly to familiarize the language usage aimed to be assessed in aviation tests. Moreover, the criteria to assess aviation English language proficiency are set by ICAO which is an international organization setting rules and regulations in the aviation sector. Therefore, ICAO and its language requirements are introduced in this chapter as well.

In order to ensure the validity of scoring in aviation English tests. The current study examines rater effects and their rating behaviors. Thus, rater effects are operationalized and a specific method to analyze rater effects, called Rasch measurement, is explained in detail. Ultimately, rater training is addressed by referring to its benefits for rater reliability in scoring.

2.2 Assessment in ESP

Research in ESP assessment does not have a long history (Swales, 1985). One of the first example of ESP testing is Certificate of Proficiency executed by the University of Cambridge Local Examinations Syndicate (UCLES) in 1913 (Park, 2015). The aim of the test was to evaluate the language proficiency of future English teachers.

Another sample for early test of ESP is English Competence Examination in 1930 by the College Entrance Examination Board in the U.S. The purpose of the test was to assess the language ability of foreign candidates in U.S. higher education context (Douglas, 2000). Although these sample tests were aimed to judge the candidates' English for academic and occupational purposes, they did not include tasks requiring target language use (TLU) domain (Park, 2015).

Due to the unrepresentativeness of ESP tests for target domains, continuous efforts have been spent on their developments. However, some researchers kept their doubt regarding ESP theories and practices (Park, 2015). For instance, Davies (2001) wonders the theoretical construct of ESP and its practices in terms of its specificity in certain contexts, which creates the problem of specific language boundaries in LSP. He gives the overlapping areas of medical and chemical English as an example. Park (2015) also states the similarity of Aviation and nautical English due to radio phraseology used in both language areas. However, Douglas (2000) claims that the considerations above lack the aspect of authenticity in ESP testing which should include tasks representing the real TLU context. In order to specify target language domains, (1) TLU context that language users will highly come across outside the tests should be identified and (2) tasks from that context should be selected (Bachman & Palmer, 1996). Focusing on these two steps will provide ESP tests with content representativeness which exemplifies the content of the target situation (Bachman, 2002). However, if the target domain is not defined well, content representativeness of tests cannot be provided easily (Park, 2015). Park (2015) gives EAP as an example since it has a broad area as TLU and it is difficult to limit its TLU to generate specific test tasks. However, it is possible to define TLU context to

select appropriate tasks for specific tests by applying needs analysis (Bachman & Palmer, 1996; Norris, Brown, Hudson, & Yoshioka, 1998).

O'sullivan (2012) states that there has been little research in modern languages as for assessment in specific purposes. The first theoretical claims based on languages for specific purposes were made by Douglas (2000). This theoretical framework is based on the assumption that language performances change with respect to their specific context and that the language occurred in that context will show unique characteristics of phonology, lexis, syntax and semantic, which differentiates it from other languages utilized in other contexts for general and other specific purposes. Therefore, Douglas (2000) looks at the concept of *authenticity* stated by Bachman (1991) to indicate the uniqueness of languages used in specific purposes. Bachman (1991) divides authenticity into two: *situational* and *interactional authenticity*. Situational authenticity points out the degree of reflection of language used for a real task in test tasks. The task performances in tests should reflect the language use domain in real life. As for interactional authenticity, the effect of an interaction among the task characteristics (linguistic, meta-cognitive and physical characteristics of performance) on task performance is taken into consideration. A test consisting of situational and interactional authenticity will allow the inferences regarding the test takers' language ability to perform in a specific language domain by observing the interaction among language ability, specific content knowledge and test tasks. These two parts of authenticity are emphasized as essential components of ESP tests by Douglas (2000) to differentiate them from English tests for general purposes. He insists that the likelihood of the way that a test taker completes the test tasks as s/he does them in real context will increase with the authenticity kept in tests tasks. He asserts that constructs intended to be measured in

ESP tests should contain background knowledge related to the specific situation to be able to make inferences about test takers' actual language abilities in certain contexts. Besides, he appends that the influence of background knowledge increases with the specialization of test contents. However, Douglas (2000) also considers the background knowledge as a confounding variable which interferes with the actual measurement of language abilities so this external factor should be restricted if possible.

The conceptualization of authenticity is criticized by Elder (2001) discussing the *distinguishability* of specific context, authenticity and the effects of non-linguistic factors. In order to detect the language usage for specific context, needs analyses (Hawkey, 1978) and corpus studies (Cheng, 2010; Flowerdew, 1997) are carried. However, an exact limit cannot be drawn between the boundaries of different language usages for specific contexts. That is an evidence for the problem of distinguishability in specific contexts. Therefore, test developers should pay attention to create a test including both tasks for test takers to use precise language for a specific context and tasks for them to show their awareness in terms of similarities and differences in language usage in different domains. In terms of authenticity, Elder (2001) claims that it is more difficult to define interactional authenticity since the cognitive aspects of the test taker while carrying out the real task should be reflected in the test tasks as well, which requires a well-designed need analysis in a specific context. Then, the validation of the test can reflect the same or similar cognitive processing of test takers for a specific language use in test tasks. Lastly, for the non-linguistic factors, Elder (2001) takes test takers' background information into account for interactional authenticity in test tasks since constructs formed in

languages for specific purposes tests consist of specific information regarding those certain contexts.

O'sullivan (2012) asserts that there has been a lack of research on the topics of test usage, localization and a theory for LSP validation (p. 78). Firstly, the researcher claims that it is the responsibility of both researchers and test takers to decide on the validation of a language test for specific purposes. Both test developers and test users should detect evidence for the tests to be used for a particular context with a specific language. Secondly, for a LSP test, localization means considering the target population and designing the test accordingly by reckoning individuals' certain language usage in that specific language domain. Lastly, O'sullivan (2012) states that there has been more focus on the practical aspects of LSP testing instead of its theoretical part and the interaction between two. Therefore, O'sullivan and Weir (2011) suggest a theoretical model for a LSP system (see Fig. 1).

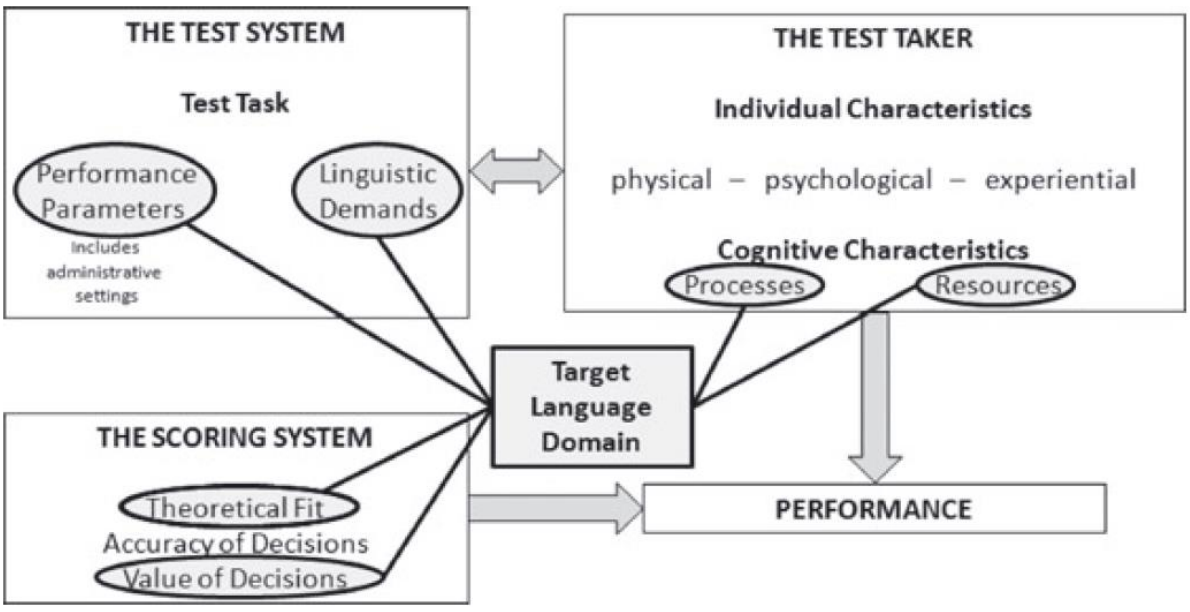


Fig. 1 A theoretical model of a LSP system
(O'sullivan & Weir, 2011)

The model takes the test taker as a central figure in the process of test development and administration. In order to make sure about the test population, both individual and cognitive characteristics should be paid attention. Firstly, individual

characteristics include physical characteristics like gender, age, physical disabilities and their duration, psychological characteristics such as memory, cognitive ability, personality and motivation and finally experiential characteristics like education and experience in the target domain. Secondly, the cognitive characteristics of the target population should be considered in detail so that test tasks can present the target domain in which test takers are forced to use similar domain specific cognitive and meta-cognitive strategies to complete the intended language tasks.

In the model, the test system answers what LSP tests test and it is divided into two as performance parameters (task settings) and language related requirements of tasks (linguistic demands). Performance parameters include the variables of timing, planning, task format and score distribution while language related requirements consist of domain specific language input and output.

Scoring system is another component of the LSP model. It comprises three parts called theoretical fit, accuracy of decisions and value of decisions (p. 82). Initially, in terms of theoretical fit, the language performance to be assessed should be in line with the constructs decided by the test developers to test domain specific language use. Besides, assessment scales, rater training and monitoring should fit with the philosophical understanding of the test. Secondly, accuracy of decisions is ensured by means of quantitative analyses in the areas like test functioning, item difficulty, bias and cut scores. Ultimately, the value of decisions is decided by checking other references for language performances. This can be teachers' formal assessments or results of other language tests so value of decisions is treated as criterion-related evidence of validity in this LSP model.

With the theoretical model of a LSP system, O'sullivan and Weir (2011) offer a theoretical base for the development of valid tests in LSP. With the presented theoretical grounding, test users can connect all the aspects of the LSP test from the development process of the test to the scoring procedures. By depending on the concept of accuracy of decisions, the current study investigates the reliability of raters' judgements in an aviation English test through both quantitative and qualitative analyses. Before, revealing the analyses directly, aviation context and the dominantly used linguistic features in it should be mentioned.

2.3 Aviation English

In the context of aviation English, there are two main interlocutors in air traffic communication: pilots and ATCs. The duty of the pilot is to fly airplanes safely without causing any accident on the air and ground while ATCs are responsible to issue instructions to give necessary clearances to avoid any incidents and to pursue air traffic. In order to communicate with ATCs, pilots use radio with distinct radio frequencies. Each ATC owns a different radio frequency and pilots can obtain these frequency numbers from different hardcopies like Flight Information Publication (FLIP) airport diagrams and low-altitude or high-altitude charts before flights. ATCs provide pilots with necessary information regarding weather phenomena like turbulence, clear runways, taxiing and air traffic conditions. Pilots need the approvals of ATCs for any maneuver to have a safe flight. ATCs can give permission to pilots based on their request or they can provide another guidance because of weather and air traffic conditions (Kaygan, 2005).

Conversations between pilots and ATCs mostly consist of commands. Pilots are the ones who mostly start the conversation to ask for information about taxing

although there are cases in which ATCs initiate the conversation to transmit significant information. Each radio transmission starts with a call sign of the other interlocutor and then stating one's own call sign so that two parts are aware of who is referring to whom. Besides, to make sure the information is received correctly, pilots and ATCs often utilize 'readback' and 'hearback' strategies. Pilots do the 'readback' by repeating the information which is gained from ATCs while ATCs listen to the repeated information to assure that it is understood correctly, which is called 'hearback' (Kim & Elder, 2015; Park, 2015). The conversation is generally completed with issuing "Roger" and "Wilco" or repeating the call sign of the correspondent to signalize that the dialog is over (Kaygan, 2005).

Aviation English is a special fragment of ESP like English for business and economy, English for science and technology (EST) under EOP (Aiguo, 2007, 2008). It is related to the aviation context in which radiotelephony communications are overly generated with the usage of both standardized phraseology and plain language by pilots, ATCs, flight crew, dispatchers, technicians and managers working in the aviation industry (ICAO, 2004). Therefore, aviation English including phraseology and plain language has certain linguistic features.

2.3.1 Phraseology

Aviation English contains repetitive, predetermined and restricted structure and vocabulary in order to avoid ambiguity both in air and ground communications. Thus, Hinrich (2008) defines ICAO phraseology as distinct formulaic phrases with minimized sound patterns and syntax lacking of pronouns, auxiliary verbs, articles and prepositions to avoid transmission of inaccurate information between pilots and ATCs. Structures used in aviation English should be simple, clear and

understandable. Words which are difficult for non-native speakers to pronounce should not be included in conversations (ICAO, 2001).

ICAO and Federal Aviation Administration (FAA) created a special phraseology for standardized aviation English telecommunication. That international standardized phraseology is shared in ICAO document 9432 called Manual of Radiotelephony (Alderson, 2011). It includes 400 linguistic items and expressions, and it is utilized internationally as the language of air (Kaygan, 2005). The aim of using standardized phraseology is to restrain from ambiguity and miscommunication which may lead to fatal accidents. For an easy understanding and impeccable interaction, both pilots and ATCs should use the phraseology to exchange necessary information so that the brief and precise information transmission will relieve ATCs to help other pilots in the frequency.

The phraseology has certain linguistic characteristics. Its unique alphabet is called ‘International Radiotelephony Spelling Alphabet’ (Park, 2015, p. 15). For instance, in order to abstain from confusing letters like B and D, interlocutors pronounce the predetermined specific words starting with the intended letter. The word ‘Bravo’ represents the letter B while ‘Delta’ means the letter D in aviation English. Apart from letters, numbers are pronounced in different ways to avoid miscommunication. As an example, number nine is pronounced as ‘nin-er’ whereas five is called ‘fife’ in order to reduce ambiguity in telecommunication (ICAO, 2001). As for delivering numbers in hundreds and thousands, each digit is pronounced separately and then ‘hundred’ or ‘thousand’ is uttered after hundreds and thousands (Park, 2015). For instance, 76200 is pronounced as ‘seven six thousand two hundred’. As presented in the short samples above, phraseology includes formulaic phrases and specific rules in utterance to prevent miscommunication in the air.

2.3.2 Plain language

Plain language is described as a genuine, natural, unprepared and uncoded language (ICAO, 2010). ICAO emphasizes that the usage of plain language should occur only at times when standardized phraseology is not sufficient to transmit the intended information. In order to pass details in unpredicted and emergency situations, plain language can be utilized in a clear, brief and precise way as the phraseology is delivered through radio (ICAO, 2010). For instance, Emery (2014) gives example extracts from a pilot and ATC who use plain language in unexpected situations:

> Pilot:
> Control. Redline 253. We have a passenger on board with suspected heart attack. We'd like to divert to the nearest available airfield. Request full medical assistance on arrival.
> or
> Air Traffic Control Officer:
> Genesis 1415. Understand you are having problems with your ailerons. Are you able to make left turns? (p. 200)

Although ICAO defines the plain language in aviation context, its explanations are not sufficient to specify the plain language so it depends on pilots and ATCs to determine the way the plain language is uttered in unusual situations (Trippe, 2018). That is why it can be an unreliable source especially in unpredicted conditions. Howard (2008) examining the authentic conversations between pilots and ATCs concludes that deviations from the standardized aviation English phraseology often result in misunderstandings in communication. Therefore, aviation personnel should restrain from using plain English as much as possible (Day, 2004) since using standardized phraseology decreases the number of misunderstandings with its precision and brevity. Another solution can be a well-defined specification of the plain language (Trippe, 2018) in aviation English in terms of complexity in structure, vocabulary and content.

Miscommunication between pilots and ATCs may happen due to the lack of understanding the intended meaning in the situations when overlapping calls, erroneous readbacks, conflict on the radio frequency and misunderstanding of flight parameters occur (Cushing, 1994). As unfortunate examples, Jones (2003) listed 35 aviation incidents ending up with 3,295 deaths between the years of 1971 and 2002. The accidents were partially due to miscommunication. As for the reason for the miscommunication in the air, Kim and Elder (2009) state that pilots and ATCs tend to use plain English in emergency situations even if radio telephony is sufficient for those times. Since usage of plain language can be complex in terms of structure and vocabulary, it can create miscommunication between interlocutors having diverse backgrounds. However, in order to prevent misunderstandings in cross-cultural communication, Kim and Elder (2009) propose the notion of interactional competence which is the ability to simplify language, avoid redundancies and paraphrase complicated sentences.

As stated above, ICAO has some regulations for the standardization of plain language to prevent miscommunication among interlocutors in the aviation sector. In order to understand the assessment criteria in aviation English tests, ICAO is introduced and its language requirements are shared with details under the next topic.

2.4 ICAO

ICAO is a particular organization of the United Nations consisting of 191 member states all around the world (Kim & Elder, 2015). The institution deals with international rules and policies regarding aviation, air safety and law. A set of language requirements were developed and published by ICAO in 2003 (Alderson, 2011). LPRs include six different levels of language skills in six areas which are

pronunciation, structure, vocabulary, fluency, comprehension, and interaction (Park, 2015). ICAO asserts that pilots and ATCs should have a certification showing their aviation English language proficiency based on the skills stated above to be able to communicate internationally (ICAO, 2004). These requirements are designed for testing plain language in an aviation English context. ICAO insists that aviation phraseology should not be tested by a language test since it is a special subject matter. It should be assessed by subject matter specialists.

In ICAO language proficiency rating scale, six language skills are included in six levels with detailed explanations. The benchmark level to get a license is operational Level 4. The overall score for a test taker is determined by the lowest score s/he is rated in any criteria (Emery, 2014). Other levels in the scale are Level 5 and Level 6 called as extended and expert levels respectively while Level 3, Level 2 and Level 1 are known as pre-operational, elementary and pre-elementary levels successively (Park, 2015). Except for obtaining Level 6 which does not require any additional retesting, test takers are retested in every three years if their proficiency level is four and they are retested in every six years if they attain Level 5 proficiency (Emery, 2014).

ICAO LPRs have been criticized by numerous researchers since it has been lacking a validation study conducted by stakeholders. Emery (2014) argues that even though ICAO language documentation includes various language abilities with side notes, the definitions of these abilities are not provided distinctly and they remain open to interpretations by stakeholders. Prinzo and Thomson (2009) also state that the level descriptors are ill defined so it is challenging to apply them while rating a test performance. Besides, Knoch (2009) questions the credibility of operational Level 4 since language descriptors for that level in the scale underrepresent language

skills in operational flying. She adds that professional experts should be included during the decision process for cutoff scores (Level 3 and 4) since language experts may not be fully aware of language usage in the target domain. McNamara (2012) also argues the invalidity of Level 4 since the responsibility for negotiation relies on just non-native speakers of English and the construct definition is lack of English as a Lingua Franca (ELF) features in the description. Therefore, further research is required to validate the ICAO rating scale and Knoch (2009) suggests that one beneficial way to validate the scale can be consulting stakeholders.

Alderson (2009) lists different licensure tests in aviation English all around the world. Some sample tests are ELPAC (English Language Proficiency for Aeronautical Communication), English Proficiency Exam for Aviators (Chile), English Proficiency Test for Aviation (EPTA) and English Proficiency Test for Airline Pilots (Japan). Forming tests and providing internationally valid licenses in aviation English is a current issue in Turkey as well. For example, PILVAK is in a test preparation period.

Overall, although ICAO releases LPRs and a rating scale for the assessment of aviation English, it does not conduct an aviation English test and does not take the responsibility to inspect the developed tests in terms of appropriateness and validation. Therefore, LPRs and the rating scale are vulnerable to different interpretations of aviation test developers all around the world (Alderson, 2010). Studies of Alderson (2009, 2010, 2011) demonstrate that the international professional standards of ICAO are not met in most of the aviation English tests. The meaningfulness, credibility and validity of current aviation English tests are questioned (Alderson, 2009, 2010). Therefore, a meticulous and close monitoring for the implementation of language evaluation policies, procedures and test qualities is

demanded. One of the ways to monitor and validate aviation English tests is conducting Rasch analysis so its details is shared under the next title.

2.5 Rasch measurement

Assessment in oral tests is a challenging process in terms of providing valid and reliable measurement. Therefore, there are particular ways to ensure validity and reliability in oral tests. For instance, rating scales with specific criteria and descriptors are introduced to raters due to the possibility of subjective judgement in oral performances. Raters are trained to apply these rating scales consistently while rating oral performances. Moreover, to make sure that raters are reliable in rating, performances with benchmarks are assigned to raters to assess. In addition to the stated practices for validity and reliability assurance, certain analyses can be utilized to detect problems of reliability and validity in speaking tests. One of them is Rasch analysis.

Rasch analysis reveals significant statistics that show both weaknesses and strengths of an assessment process. Rasch hypothesis assumes that there is a single scope of measurement for test takers' ability and task difficulty (McNamara, 1996). Thus, Rasch models illustrate assessment parameters (e.g., test taker ability, item difficulty and rater severity) in *logits* which are the transferred mathematical representations of observed scores. The models are set to calibrate each variable independently with the aim of placing the raw scores on a logit scale (Barkaoui, 2013). The scale is an interval scale in order to demonstrate relationships among variables.

Diverse models have been generated depending on Rasch model. Multi-faceted Rasch model (MFRM) is an extended model of Rasch and it was created by Linacre

(1989). It is also called many facet Rasch measurement (MFRM) (Eckes, 2009). Its distinction is in two areas. Firstly, more than two facets can be analyzed at the same time. Secondly, the data selected for analyses can be polytomous in addition to dichotomous.

MFRM has been increasingly used for assessment in the areas of psychology, education, medicine and language (Eckes, 2009). It provides the identification and measurement of irrelevant variables which affect test takers' scores systematically. These variables are called *facets* in MFRM (Barkaoui, 2013). Facets which influence the results can be due to raters, rating scales or criteria and task types (Küçük, 2017). The quantity of the impact caused by facets on the measurements can be estimated thanks to MFRM. With the estimation, MFRM compares facets on a ruler. All the facets (test takers, tasks, raters and rating criteria) are placed on a true interval scale in logits, called Wright map (Wilson & Draney, 2000). Wright map facilitates the comparison among different facets. It illustrates the interaction of facets by revealing statistics for test takers' abilities, raters' severity/leniency and difficulty in tasks and criteria. On Wright map (see Fig. 2), the positive values above zero show higher test taker ability, rater severity and more difficulty in tasks and criteria whereas the negative values below zero illustrate the reverse.

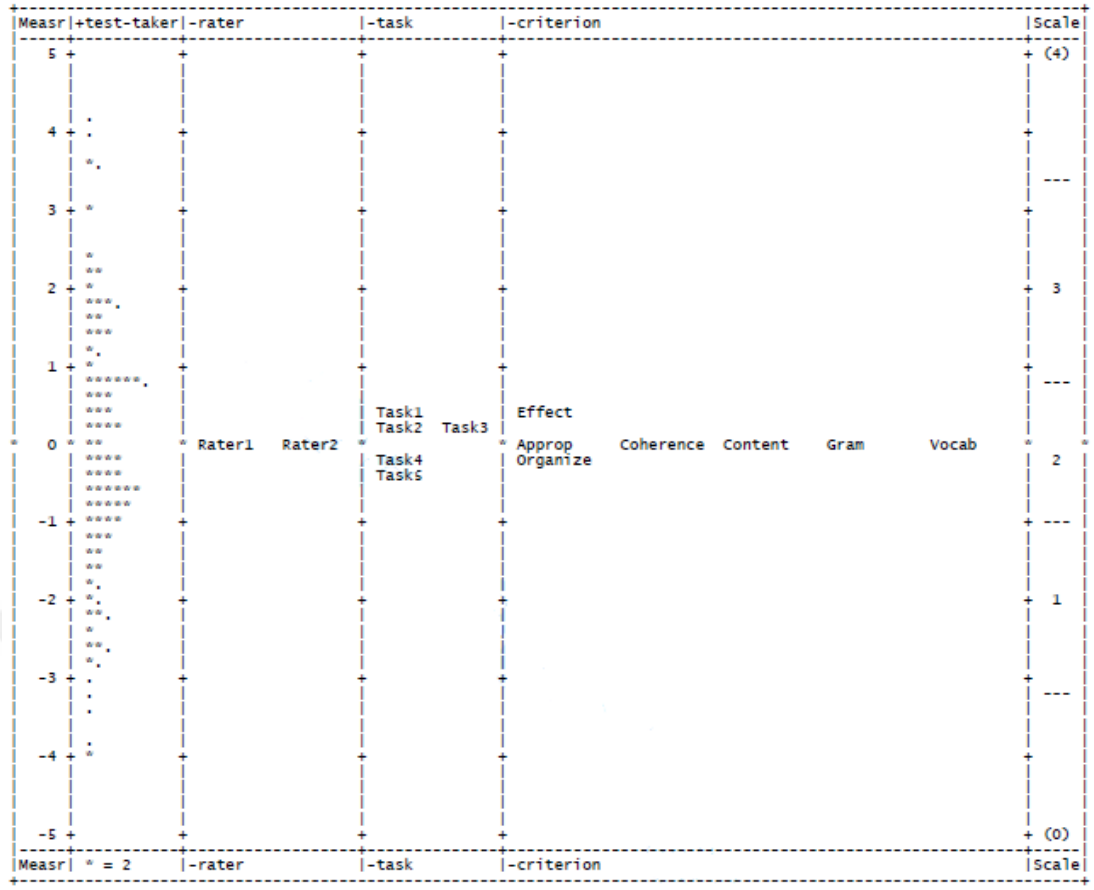


Fig. 2 A sample Wright map
(Barkaoui, 2013, p. 46)

To calculate estimation by depending on a frame of reference in MFRM, there should be enough interconnections among all components of facets (Linacre, 1996). Barkaoui (2013) states that all raters' evaluation of the same short parts in test takers' performances will be enough to link the facets for MFRM. In order to carry out calculations for MFRM, the computer program Facets was developed by Linacre (2007). The program uses the scores that are attained by raters to test takers' performances so as to conduct statistical estimations for each variable. FACET displays values including standard error (SE) and fit statistics for each component of facets to prove the reliability of the estimations and validity of the measures (Barkaoui, 2013). The program can also be utilized to investigate task and criterion difficulty, rater severity/leniency, rater consistency, bias (interaction) analysis, rater training process and rating scales. Furthermore, interactions between different facets

can be found by checking the standardized residuals, which are the difference between the observed score and the expected score in the model, and unusual patterns in interactions between facets can be detected with bias (interaction) analysis as well (Eckes, 2005).

Facets provides statistical data for both group level and individual level analyses regarding fair average, SE, separation index, reliability of separation and infit and outfit mean-square for each facet (Barkaoui, 2013):

- Fair averages are the scores that are free from rater effects so they illustrate possible genuine proficiency levels of examinees and they are helpful for fair comparisons between scores (Eckes, 2009).
- SE indicates the ambiguity in the estimations.
- Separation index shows how many levels occur in facets. For instance, it can reveal the number of severity levels for raters and the ability levels for test takers.
- Reliability of separation demonstrates to what extent the analyses distinguish the levels in each facet truly.
- Infit and outfit mean-square statistics indicate the amount of inconsistency in each element of facets (a specific rater or test taker) with respect to the degree of variability in each facet it belongs to (within the raters or test takers). While an accumulation of unexpected scores affects infit mean-square values (Myford & Wolfe, 2003), outliers in the data influence outfit mean-square more (Linacre & Wright, 2004). Observed and expected scores from the model created in FACET are compared and if they correspond to each other, the fit statistics get closer to the value of 1.0 which is considered as an ideal infit and outfit mean-square statistic. Moreover, the value under 0.5 signals

overfit, which means the data is too expected, whilst values over 1.5 shows misfit, which means data is unpredictable and inconsistent. That is to say, the values between the ranges of 0.5 and 1.5 (Linacre, 2002) are acceptable. The closer the values are to 1.0, the better results the statistics reveal. In a similar vein, McNamara (1996) suggests a narrower range with the control limit between 0.70/0.75 and 1.30 for high-stakes testing.

For better accuracy in measurement, a larger number of raters can rate the same examinees. This leads to more precision in model estimations for raters, test takers and tasks. Moreover, for the highest precision in model estimations, all test takers should be rated for all tasks by all raters. This process is called "complete" or "fully crossed rating design" (Eckes, 2009, p. 39). It should be stated that a complete crossed rating design is followed for the current thesis as well. In addition to that, this study utilizes Rasch analysis specifically to investigate rater effects in an oral performance evaluation process so rater effects should be operationalized before conducting the analyses.

2.6 Rater effects

In performance-based assessment, test takers are expected to create a performance instead of selecting correct answers provided in tests. Therefore, they need to construct a related response for each question. After these tests, the interpretation, assessment and rating of the responses require thorough investigation to represent objective judgements. Thus, psychometric calculations can be beneficial to ensure the quality of rater-mediated assessment (Eckes, 2009).

One of the major problems faced by testing practitioners is the construct irrelevant rater variability in rater-mediated performance assessment (Eckes, 2009).

Rater variability results in construct irrelevant deviance in test takers scores which, in turn, threatens the truthfulness and validity of assessment procedure. Rater effects, errors and bias are related terms for rater variability in assessment. In this thesis, rater effects are utilized in order to indicate rater variability in ratings.

Rater effects are systematic deviations in scores attained by raters (Myford & Wolfe, 2003) since they are irrelevant variables to the constructs that are aimed to be assessed (Weir, 2005). Therefore, they jeopardize the validity of the assessment. To retain the validity in the assessment procedure, Facets is a beneficial statistical tool to detect systematic rater effects including rater severity/leniency, inconsistency, central tendency, halo effect and rater bias (Myford & Wolfe, 2004).

When raters rate certain tasks or criteria more severely or leniently, they may show *severity/leniency effect* (Knoch et al., 2007). However, Eckes (2009) does not see the difference in rater severity levels as problematic as long as raters are consistent in their rating behavior. If raters are not consistent in their ratings, this can be an evidence for *inconsistency effect* (Knoch et al., 2007) which is called randomness by Myford and Wolfe (2003). Additionally, if raters are remarkably consistent in their ratings by overusing the middle categories and underusing the extreme categories in rating scales, this may indicate *central tendency effect*. Moreover, raters may assign similar scores to a test taker on various criteria although s/he shows distinct levels of abilities for different criteria. This situation indicates a *halo effect*. In these circumstances, the rating criteria can be revised for clarification or raters may need more training to distinguish criteria in rating scales.

The unusual patterns which are consistent deviations from the expected measures in the Rasch model can be identified with interaction (bias) analysis. Such

deviations in facet elements are indications of differential facet functioning (Du, Wright & Brown, 1996). Raters may rate certain criteria more severely than others. If there is a consistent pattern in ratings in terms of severity for specific criteria in rating scales, raters may show *rater bias* against those criteria. In addition to criteria, bias analyses can be conducted to investigate raters' severity levels towards test takers, tasks and across time.

As indicated above, interaction analysis can detect differential rater functioning, i.e. rater bias (Myford & Wolfe, 2003) in raters' interactions with test takers. It tests the hypothesis which claims the absence of bias except for measurement error. In MFRM, rater bias statistics are illustrated by tables and charts to identify unusual rater behavior easily and the statistics for bias analysis are given in t-values. The control limits are between t = 2.0 and t = -2.0 (Eckes, 2009). The values above 2.0 and under -2.0 should be investigated for any rater bias.

In terms of interrater reliability, Eckes (2009) mentions the "agreement-accuracy paradox" (p. 8). He claims that high consistency among raters (high reliability) does not guarantee the accuracy of ratings. Thus, high rater agreement may result in wrong judgements regarding scores. However, with a measurement approach, this paradox can be resolved. Eckes (2009) indicates that MFRM is a beneficial tool dealing with the paradox above thanks to its detailed statistical analyses to detect similarities and differences in ratings. For instance, in order to assure the quality of rater mediated assessment procedures, Eckes (2005) utilized Facets to detect rater effects in the writing and speaking parts of Test of German as a Foreign Language (TestDaF). He found out that raters differed in severity but they were internally consistent in scoring the written and spoken output of test takers. However, interactions between the facets of 'Rater X Criterion' and 'Rater X Task'

revealed nearly 37 % and 16% unexpected variance from the model expectations successively. These deviations from the model exhibited the undesirable rater variability in the assessment procedure.

The stated variables in raters' rating behaviors can be prevented or decreased with the help of rater trainings. Therefore, the aim of the rater trainings and the expectancy from them are referred under the next subtopic.

2.7 Rater training

The purpose of rating training is to help raters understand the construct being assessed and, then, to calculate statistics for both inter and intra rater reliability so as to make sure the absence of rater error in scoring (Eckes, 2009). McNamara (1996) stresses that the significance of rater training is keeping the raters internally consistent in their own scoring. In rater training procedure, the construct aimed to be measured, criteria and categories in rating scales, levels of task difficulties and the performance levels which tests aim to measure should be introduced in detail in order to calibrate raters' scoring and to reduce rater effects on ratings (Eckes, 2009).

McNamara (1996) thinks that the aim of the rater training is to retain internal rater consistency but there can always be variability among raters due to their distinct characteristics. Hence, Elder, Knoch, Barkhuizen, and Randow (2005) recommend that keeping within rater consistency stable can be an effective way in rater training programs. Likewise, Knoch et al. (2007) state that rater training is beneficial to increase within-rater consistency and to reduce differences in rater severity and bias.

As stated above, it can be unrealistic to keep the raters internally consistent but with MFRM statistics, raters can be monitored and given feedback regarding their own consistency in ratings. Thanks to the individualized feedback provided

with statistics in rater training programs, within rater consistency can be increased. Eckes (2009) suggests following components for rater feedback: (a) measures of rater severity or leniency, (b) statistics of infit and outfit mean square for within rater consistency, (c) the usage rate in rating scale categories and (d) rater bias. In a similar vein, the results of this thesis can be utilized as individualized feedback for each ETAP rater since all components mentioned above are included in the statistical analyses of the present study.

Raters' inferences should be based on criteria in rating scales. However, there can be some scale extensions which are certain criteria not indicated in rating scales (Kuiken & Vedder, 2014). Thus, it is significant to detect whether raters' rating behaviors correspond to the aims of the test or not. Kuiken and Vedder (2014) think that raters may pursue certain rating characteristics and there can be some reasons in their rating behaviors. In order to investigate different rating behaviors, raters can be asked to dictate why they assign certain scores to certain performances. With their oral reports, the use of criteria and the degree of importance given to each criterion can be found out. That is to say, qualitative analyses are required to investigate the way raters use rating scales and their cognition.

Barkaoui (2013) suggests that other data analyses should accompany MFRM analyses for a depth investigation in assessment process. Qualitative data can be collected with the help of oral protocols, interviews and observation so as to understand the reasons behind the MFRM statistics. A study with mixed methods including the analyses of qualitative data before, during and after the MFRM analyses is beneficial to examine the assessment process and to form valid arguments for it. Kim (2015) also states that to have a better understanding for raters' rating behaviors, qualitative studies can be conducted to examine raters' decision process.

The way they use rating scales, their interpretation of descriptors in rating scales, their attentions across tasks and proficiency levels of test takers can be analyzed thanks to verbal reports provided by raters themselves. For instance, Knoch et al. (2007) compared the effectiveness of online rater training with the traditional face to face rater training in writing assessment. Sixteen raters were divided into two groups. Seventy scripts were assigned to them before the training sessions. Then, two groups scored 15 scripts during their training sessions and got individualized feedback upon their ratings. After the training, both groups assigned scores to the same 70 scripts again. Raters were asked to answer the questionnaires and to participate in interviews regarding their different way of training. Rater severity, internal consistency, central tendency, halo effect and rater bias were compared statistically in both groups and MFRM results showed the overall effectiveness in both training types. Both online and face to face rater training were helpful to reduce severity, central tendency and inaccuracy in ratings. However, qualitative results demonstrated variability in favoring different training types.

Bogorevich (2018) also utilized a mixed methods approach including facet analysis and thematic coding in order to investigate the differences between native and non-native English-speaking raters in their evaluation of oral performances. In her thesis, she examined raters' cognition and perception regarding the process of their rating. She analyzed think-aloud reports and raters' interviews. After the thematic analyses, it is found that there were diverse strategies of raters in terms of listening to and grading the performances. Besides, raters' perceived severity and category importance were examined in depth. Additionally, non-rubric criteria were coded for both delivery and theme development. As for delivery, voice quality and accent familiarity were found as non-rubric criteria. Likewise, finished and

unfinished answers, organization, reading the prompt, making a decision and the quality of the thoughts were identified as irrelevant criteria for theme development. Similarly, Ang-Aw and Chuen Meng Goh (2011) benefit from descriptive statistics and oral report analyses to investigate rater differences in a high stakes test in Singapore. Firstly, it is found that raters differ in their focuses on factors to be assessed. Secondly, they perceive the targeted oral constructs differently. Thirdly, the way they interpret the oral performances and scores change. Lastly, raters follow diverse approaches while evaluating the performances. Overall, Differences in rater judgements were found in spite of provided rater training.

With the stated significance of using a mixed methods study to investigate rater reliability and rating behaviors before, during or after rater training sessions, both quantitative and qualitative analyses are conducted in the current study to reach an extensive investigation upon rater effects and their foci in an oral assessment process of an aviation test.

2.8 Conclusion

In the aviation industry, safety comes first and one of the most significant issues in air safety is communication. More than 1,000 individuals lost their lives in only three aviation accidents due to miscommunication (Ripley & Finch, 2004). Therefore, communication skills of ATCs and pilots flying internationally play huge roles in air safety. They need to understand English aviation phraseology and plain language in order to communicate flawlessly. ICAO is the supreme authority which proposes language requirements for aviation English tests (ELPAC, IELTS and TEA) in order to make sure English proficiency levels of pilots and ATCs in international flights (Alderson, 2009). However, setting language requirements may not be enough for

language standardization among internationally working pilots and ATCs because of the subjective performance assessment in oral tests. Rater training for rating calibration is needed for test takers to gain fair scores out of these oral tests but raters may show different rating behaviors even after taking rater training. For instance, raters can pay more attention to certain descriptors based on test takers' proficiency levels (Pollitt & Murray, 1996) and depend on their own interpretations to assess oral performances (May, 2009). Therefore, the patterns in raters' rating behaviors need to be investigated to understand the rating process better (Kim, 2015). As for assessing oral parts of aviation English tests, raters have even bigger responsibility upon them. They need to be precise and consistent in their ratings for air safety since certification of pilots and ATCs for international flights depends partly on raters' evaluations. It seems particularly difficult to assess performances in aviation tests for raters due to the interactive and particular features of aviation language. Therefore, special rater training is required for aviation test raters to get accustomed to the aviation context and language features.

To wrap up, the background information related to the current research is provided by referring to the details of assessment in ESP, aviation English, ICAO, Rasch measurement, rater effects and rater training in this chapter. Next chapter presents the methodology for this study by stating the research purposes, participants, data collection procedure and data analysis.

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter presents the research methods and procedures followed for the online rater training period and for the current research. The purpose of the research, utilized instruments, rater training and data collection procedures are explained in depth. The chapter ultimately ends with the research questions and detailed explanations of data analyses used to examine those questions.

3.2 The purpose of the research

Unfair scores can be assigned to examinees in performance-based speaking tests due to the inconsistencies in raters' rating behaviors. The variables in raters' scoring may lead to undesirable outcomes in aviation sector since miscommunication may emerge among the personnel having low level of English proficiency. With the realization of rater behavior effects on assessment, the purpose of this study is to investigate whether the raters trained online express rater effects while evaluating the speaking performances of the pilots and ATCs taking ETAP. Especially, whether ETAP raters show the effects of leniency/severity, consistency, central tendency, and rater bias towards criteria in the scales are investigated. Moreover, the focuses of these raters while rating are probed by analyzing ETAP raters' verbal reports.

The following research questions were formulated for investigation in this study.

1. Do ETAP raters differ in their scoring in terms of their individual level of severity and leniency when assessing ETAP test takers' oral performances?
2. How consistently do ETAP raters rate ETAP test takers' oral performances?
3. Do ETAP raters show any central tendency effect due to individual indistinguishability of score levels in ETAP rating scale?
4. Do ETAP raters show any rater bias towards any criteria in the rating scale? If yes, for which criteria do the raters behave more severely or leniently?
5. What do ETAP raters focus while rating oral performances?

3.3 Participants

Convenience sampling (Merriam & Tisdell, 2015) was used to select participants due to their availability for this study. By signing a legal document proposed by PILVAK, ETAP raters agreed upon that the data including grading and voice recordings could be utilized for any scientific purposes for the field of language education. Since the researcher of this study was an ETAP rater, it was easy to reach the raters and take their oral consent to analyze the data for the research. Then, all the data were provided to the researcher by PILVAK via emails. In total, there were eight raters who participated in the rater training sessions. However, four of them did not systematically send their final scores and oral recordings. Since the recordings were crucial for the qualitative part of this study, half of the participants were excluded from the research. In total, four ETAP raters including the researcher herself participated in this study and their qualitative and quantitative data based on

eight pilots' oral performances were utilized for the research purposes. Apart from this, pseudo names (Ege, Olcay, Deniz and Bahar) were assigned to the participants for this study in order to protect confidentiality. As for the raters' personal information, since effects of raters' background, personality characteristics and beliefs (Myford & Wolfe, 2003) are not the scope of this study, detailed background information regarding the raters was not collected. However, it is known that the participants have taught general English before and some of them are experienced in teaching aviation English as instructors.

3.4 Data collection instruments

3.4.1 ETAP

In this part, ETAP is introduced briefly with the help of its validation document (Ünaldı, 2019). Anyone who is interested in ETAP can consult that manual. Below, the purpose of the test, test tasks, rating scale and scoring process are shared in order.

ETAP is an ESP test created by PILVAK thanks to the collaboration with Boğaziçi University Lifelong Learning Centre (BUYEM). PILVAK is a non-profit organization and does not possess any commercial interest in conducting ETAP except for supporting pilots and ATCs to obtain a valid certification in aviation English. The aim of the test is to evaluate English proficiency levels of pilot candidates, commercial pilots and ATCs in a job-related context. The test is prepared in line with ICAO standards and language requirements. Although the test is developed in Turkey, it is assumed that its design and content do not include any reference for cultural and first language background so it is appropriate for international candidates from diverse backgrounds. Moreover, ETAP does not assess the correct use of aeronautical phraseology, professional knowledge and experience

in aviation but test takers should be aware of standardized phraseology and basic performances in aviation because test tasks include routine and non-routine aviation context relevant situations. The test tasks do not include general personal and job-related answers that can be memorized beforehand. Topics of tasks are context bound but unpredictable so it is aimed for test takers to improve their language skills while studying for the test.

ETAP includes listening and listening into speaking tasks in order to keep the authenticity of the tests by trying to balance contextual relevance and cognitive load of the test tasks. As for authenticity, interlocutors ask some follow up questions to the test takers after the listening tasks in order to sustain the nature of interaction. However, there is no second chance for listening to the audios in the test although participants in aeronautical communication can ask for clearance and repetitions. Since this is a test, it is difficult to grade the gradual process of test takers if they form better linguistic outputs after listening more than once. Besides, in order to summarize the listening text, candidates are supposed to select the important information and process it in their working memories so it is significant to evaluate the quality of their naturally occurring understanding. As for background noise, since it is not objectively known how much of it would be normal in aeronautical communication, the use of background noise in the test is avoided.

ETAP consists of two parts. In the first part, there are three listening tasks for comprehension. Short listening questions are provided to test takers on a paper and time for reading the questions before the tasks is allocated for the participants. The first part of the test aims to assess the listening skill at Level 3 and 4. In the second part of the test, there are three integrated listening into speaking tasks and one picture description task. The spoken responses of test takers are recorded for later

assessment by raters. The second part aims to assess the language proficiency levels from Level 4 to 6. There are detailed explanations and sample questions for each task in the ETAP validation document (Ünaldı, 2019). Since this thesis focuses on rater effects in speaking assessment, Tasks 4, 5, 6 and 7 in ETAP are explained briefly.

In Task 4, five different non-routine situations are delivered to the test takers who are supposed to ask three questions to get more detailed information about the situation. Test takers are expected to ask fifteen questions in total. Their questions should be context relevant and include correct use of structure and related vocabulary. Task 4 aims to assess proficiency Level 4.

In Task 5, three pictures showing a flow of an incident are provided to test takers who are supposed to talk about the event illustrated in the pictures. After narrating the event, test takers are asked three follow up questions regarding the possible reasons and solutions of the problem. Describing an action in the present or past, providing explanation, source of the problem and solution are aimed as linguistic output. The task aims to assess proficiency Level 4 and 5.

In Task 6, an aviation related event is narrated. Firstly, test takers are expected to summarize what they hear from the recording and then they should answer five or six follow up questions. The comprehension, paraphrasing, definition, explanation and narration abilities are aimed to be assessed. Test takers need to describe the process and respond to the questions depending on their background knowledge. They can give reasons, provide solutions, and make predictions by relating the follow up questions with the given event. The task aims to assess proficiency Level 5.

In Task 7, test takers listen to a completely authentic recording about an aviation incident. The audio is not simplified and can include common and unknown vocabulary items. Participants are asked to summarize the audio and then they need to answer five or six follow up questions. The reason of the incident is not provided in the listening text so in the follow up questions, the reasons of the event and precautions to prevent it are requested as answers. In the follow up question part, test takers are also expected to speculate on the provided event and share their own ideas regarding it. The task aims to assess proficiency Level 6 in which extended speech, understanding distinct words, idiomatic expressions and cultural points, advanced language usage in common and uncommon topics are demanded.

The outputs of test takers in different tasks cannot be evidence for all constructs in descriptors of ICAO rating scale for which a revision is suggested (Alderson, 2009). For instance, it is not proper to use picture description tasks to assess language comprehension skills of a test taker. Therefore, different tasks may require more specialized descriptors in rating scale rubrics. As such, more detailed rating scales (see in Appendix A) for each ETAP task were developed by extending ICAO rating scale. In ETAP, there are three distinct rating scales. One is for Task 4, one is for Task 5 and the last one is for Tasks 6 and 7. In the rating scales, different descriptors take place to assess abilities in different tasks. In Task 4, test takers' performances of language usage in structure, vocabulary and comprehension are rated whilst in Task 5, the usage of pronunciation, structure, vocabulary, fluency, interaction and content are assessed. In Task 6 and 7, test takers' performances in pronunciation, structure, vocabulary, fluency, comprehension, interaction and content are evaluated by raters. Since tasks require diverse and distinct linguistic abilities, the construct descriptions slightly change in each ETAP rating scale.

In terms of scoring, the first three listening tasks are scored objectively because test takers should answer the written questions on the test paper. In the speaking part of the test, the test taker interacts with the interlocutor and s/he is recorded to be assessed later by two different raters. Raters are unaware of the scores attained by the other rater for the same test taker. Before actual scoring, in order to get standardized, raters assess sample recordings and receive immediate feedback on their rating performance. Besides, raters can be assigned pre-scored sample recordings among their actual ratings so that their rating reliability can be monitored continuously. If the difference between the scores given by two raters is one band in the final score, a third rater scores the same test taker's performance. The final score is decided based on two similar scores. If three different scores appear at the end of three scoring, the middle score is approved as a final score. The total scoring scheme is shared below (see Table 1) and the contraction 'T' means 'task' in the table.

Table 1. Scoring Scheme for ETAP

Task 1 and 2	1 point for each correct answer
Task 3	1,5 points for each correct answer
Pronunciation	(T5+T6+T7):3
Structure	(T4+T5+T6+T7):4
Vocabulary	(T5+T6+T7):3
Fluency	(T5+T6+T7):3
Comprehension	((T1-T3 average) +T6+T7):3
Interaction/Content	(T5+T6+T7):3

It should be noted that in the rubric of Task 4, structure and vocabulary are assessed together under the criterion of Structure/Vocabulary and the score given to that criterion is added with the scores of criterion structure in Task 5, 6 and 7. Then, they are divided into four to obtain the average score.

3.4.2 Recordings of pilot performances

Examinees participated in oral tests in a test center owned by PILVAK. Their oral performances were recorded by the interlocutor processing the exam. All recordings of pilot performances were shared with ETAP raters by PILVAK via drive. For this study, eight pilots' performances of different pilots were assessed by four ETAP raters. Each performance recording was around forty minutes, started with greetings and followed the task order.

3.4.3 Scores and verbal reports

Each ETAP writer was supposed to assess the pilots' oral performances, keep their verbal reports and send them to a PILVAK secretary via email until the due dates decided beforehand. For the purpose of archiving the assigned scores, the institution shared Excel sheets named for each pilot with ETAP raters. Raters entered their scores into the Excel documents and the researcher piled up the Excel documents provided by PILVAK with the aim of quantitative data analysis for the thesis.

After assigning scores to the pilots' oral performances, ETAP raters stated their judgements in their oral recordings. They explained the reason why they assigned specific scores to the performances. In the recordings, they generally refer to each task successively by focusing on the descriptors in the criteria. Four raters recorded 55 oral reports in total. As shown in Table 2, the differences in the numbers of recordings per rater are due to the way the raters kept the reports. That is to say, while Deniz and Olcay mostly preferred to record their evaluations task by task, Ege and Bahar kept the recordings by mentioning every task separately in one recording for per performance. The length of the reports ranged from 29 seconds to 6.59 minutes and the recordings were 150 minutes long in total. In the recordings, raters

spoke Turkish, however they did code-switching so they used English from time to time.

Table 2. Descriptors for ETAP Raters' Verbal Reports

	The number of recordings	The longest recording	The shortest recording	The total duration of recordings
Deniz	17	6.59 min.	1.40 min.	72.35 min.
Ege	8	3.15 min.	0.50 min.	16.54 min.
Olcay	22	2.47 min.	0.29 min.	34.26 min.
Bahar	8	4.50 min.	2.06 min.	26.05 min.
Total	55	-	-	150 min.

The oral reports were obtained by emails and piled up to be transcribed and analyzed. The transcription of the verbal reports was carried out by two researchers and the analysis was conducted by the researcher of this thesis. The codes and themes formed out of the data were double checked by the researcher herself.

3.5 Procedures

ETAP rating training sessions lasted eight weeks in total. Participants attended online meetings for at least weekly three hours through Skype. The meeting dates and times were arranged according to the availability of the participants in order to ensure their regular attendance in each online session. The online sessions on Skype were recorded so that the participants could watch them whenever they wanted. For the first three weeks of the online rater training, the rating scale following ICAO standards was introduced and participants were asked to evaluate 22 short oral performances which were rated before for training purposes. These short speaking performances had standardized benchmarks and were assigned to ETAP raters to evaluate for the purpose of rating calibration. Later, the given scores for the

assignments were discussed in the online sessions. In the last five weeks, the focuses of the online rater training were ETAP exam items, the test specifications and ETAP rating scale. After introducing ETAP in general, speaking performances of eight pilots were assigned to ETAP raters both for a deeper examination of the test and for rating calibration. The attributed scores to the performances of ETAP test takers by ETAP raters were discussed in the online sessions. After each rating, verbal reports regarding raters' judgements about scores were demanded as a requirement for future research. Before each online session, the attributed scores and oral reports were obtained from the participants via email.

3.6 Data analysis

The procedure for the data analysis is reported separately for each research question. The rater behavior was investigated via multi-faceted Rasch measurement (Linacre, 1989). To analyze quantitative data, the limited version of Rasch software Facets called *Minifac* (Version No. 3.83.3) by Mike Linacre (https://www.winsteps.com/minifac.htm) was utilized. As for the qualitative part of the study, the verbal reports, which are participants' oral recordings to present their ideas during (think-aloud) or after (retrospective) scoring tasks (Heigham & Croker, 2009), were transcribed and analyzed via the software NVivo 12.

3.6.1 Do ETAP raters differ in their scoring in terms of their individual level of severity and leniency when assessing ETAP test takers' oral performances**?**

Rater severity is defined as the raters' tendency to rate the performances lower than other raters do for the same ratees (Myford & Wolfe, 2004). Severe raters underestimate the performance level of the test takers and assign lower scores than

other raters do for the same test takers. On the contrary, lenient raters tend to overestimate ratees' performances and assign more score than other raters do for the same ratees. In order to answer the first question, the Wright map (Wilson & Draney, 2000) in the Facets output was checked to see the severity levels of the raters in order. The most severe rater appears on the top of the map while the most lenient one is placed on the bottom. Then, raters' measurement report (see Fig. 3) in Facets was scanned. In the table, measures for rater severity, separation and reliability statistics (Küçük, 2017) were utilized.

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	N Raters
617	144	4.28	4.16	-1.47	.13	.94	-.4	1.25	1.6	.81	.69	.76	36.1	42.4	3 Olcay
660	144	4.58	4.54	-2.17	.13	.87	-1.1	.79	-1.6	1.23	.81	.78	46.8	45.6	1 Deniz
665	144	4.62	4.59	-2.25	.13	1.02	.2	.87	-.9	1.10	.83	.78	42.4	45.7	4 Bahar
711	144	4.94	5.03	-3.04	.13	1.09	.7	.97	-.1	.91	.76	.78	34.0	42.7	2 Ege
663.3	144.0	4.61	4.58	-2.23	.13	.98	-.2	.97	-.3		.77				Mean (Count: 4)
33.3	.0	.23	.31	.56	.00	.08	.7	.17	1.2		.05				S.D. (Population)
38.4	.0	.27	.36	.64	.00	.10	.8	.20	1.4		.06				S.D. (Sample)

Model, Populn: RMSE .13 Adj (True) S.D. .54 Separation 4.17 Strata 5.90 Reliability (not inter-rater) .95
Model, Sample: RMSE .13 Adj (True) S.D. .63 Separation 4.85 Strata 6.80 Reliability (not inter-rater) .96

Fig. 3 ETAP raters' measurement report

The measure column in the table shows the severity levels of the raters. There is a reverse relationship between the measure and the total score of the test takers since the rater with higher total score tends to give more score to the performances so s/he is the least severe and the most lenient one. Therefore, the most severe rater is located on the top of the measure column. Moreover, separation and reliability statistics indicate that raters are reliably different from one another. Reliability statistics range from 0 to 1 which indicates perfect separation, meaning significant difference in severity levels (Lumley & McNamara, 1995). Separation statistics show how many distinguishable severity levels of different raters there are. Therefore, the rater separation statistic should be close to 1.0 (Eckes, 2009) and the reliability

statistic should be close to 0 (Myford & Wolfe, 2004; Yılmaz, 2017) in order to claim that the raters have similar severity levels so they are interchangeable.

3.6.2 How consistently do ETAP raters rate ETAP test takers' oral performances?

Raters may behave inconsistently while evaluating performances. To interrogate the individual consistency in ratings, infit and outfit mean square (MnSq) statistics (see Fig. 3) which reveal raters' reliability were examined. Infit and outfit MnSq values report whether the raters are consistent in evaluating students' performances or not. If the infit and outfit MnSq statistics fall in the range between 0.5 and 1.5 (Linacre, 2002), the raters on average are reliable, which means that they are self-consistent in terms of rating ratees' oral performances. The value of 1.5 and above indicates misfit, i.e. the inconsistency and unpredictability of raters while the value of 0.5 and below points out overfit, i.e. the over-predictability of raters' assessment (Lunz, Stahl, Wright & Linacre, 1989), which means that raters give similar scores to performances. Apart from infit and outfit MnSq, z-standardized (ZStd) values in the Facets table can be investigated. ZStd values should fall between -2 and +2 (Boone, Staver, & Yale, 2014). However, If the sample size is small. ZStd can be easily inflated or deflated. In that case, infit and outfit MnSq should be relied upon. Besides, ZStd is ignored when infit and outfit MnSq values are between acceptable ranges (Boone et al., 2014).

3.6.3 Do ETAP raters show any central tendency effect due to individual indistinguishability of score levels in ETAP rating scale?

Raters may demonstrate a central tendency effect in their rating behavior by avoiding extreme categories in rating scales and overusing middle categories. (Myford &

Wolfe, 2004). That is to say, raters may not be able to differentiate the scale categories and they tend to assign middle categories to ratees' performances at most. To examine central tendency effect, rater infit and outfit MnSq values, a partial credit model for raters and test takers' separation statistics were analyzed (Küçük, 2017). Infit and outfit MnSq show whether the measures are affected by other variables or not in order to see the reliability of these measurements. As stated, they indicate whether the raters are reliable or not. The reliable ranges for the statistics of infit and outfit MnSq should be between 0.5 and 1.5 for Facets analysis (Boone et al., 2014). Any value under 0.5 is overfitting, which means that raters are too predictable. In order words, they tend to overuse middle categories in rating scales (Knoch, 2007). However, Myford and Wolfe (2004) states that this is not always the case. To make sure presence of central tendency effect, *Hybrid Model #2* and *Hybrid Model #3* (Myford & Wolfe, 2003, p. 414) can be utilized to understand raters' individual use of a single trait in a rating scale. In ETAP raters' case, the formulas '?,?,#' and '#,?,#' were utilized successively to apply Hybrid Model #2 and Hybrid Model #3 into Facets program. For these models, partial credit models which illustrate certain facets in detail were applied for raters and criteria. Furthermore, test takers' separation statistic indicates the discrimination power of the measurement so it shows the distinguishable ability of the test takers by raters. A high value means that test takers are significantly different in terms of their oral performance scores attributed by the raters, which in turn reveals evidence against the presence of central tendency effect (Sudweeks, Reeve, & Bradshaw, 2005).

3.6.4 Do ETAP raters show any rater bias towards any criteria in the rating scale? If yes, for which criteria do the raters behave more severely or leniently?

The variability of raters with respect to the other facets was examined through bias analysis. The severity and leniency degrees of raters towards specific variables were measured in order to understand causes of rater bias to improve rater training and rating scale (Schaefer, 2008). Individual rating patterns and their effects on performance assessment can be obtained by bias analysis in Facets. If a rater rates a specific criterion too severely or leniently, s/he may be considered having bias towards that criterion in the scale (Knoch et al., 2007). By conducting bias analysis, McNamara (1996) found out the tendency of raters towards grammatical accuracy of candidates in the Occupational English Test (OET) consisting of a communicative feature. Likewise, in online ETAP rater training sessions, the researcher witnessed that certain criteria were overestimated by raters. They might rate specific criteria more severely than others. For instance, it was assumed that the criterion 'structure' could be assessed more severely than the criterion 'comprehension'. Thus, to identify raters who are biased for certain criteria, a 'Rater x Criteria' bias/interaction analysis was conducted and t-values which are greater than +2 (for severity) and smaller than -2 (for leniency) were checked to find out significant rater bias (McNamara, 1996; Lumley & McNamara, 1995; Knoch et al., 2007). Rater bias patterns were identified and a pairwise comparison was applied to observe the differences among raters.

3.6.5 What do ETAP raters focus while rating oral performances?

To understand rater behaviors better, a mixed methods approach providing both quantitative and qualitative data analyses was employed (Shirazi, 2019). The underlying reasons for rater behaviors were investigated with the help of qualitative

data analysis (Schaefer, 2008). Thus, ETAP raters' verbal reports were transcribed and analyzed in order to examine cognitive processes and behaviors of raters (Kasper, 1998). To make interpretations, specific categories and themes were generated through NVivo 12. The research question, ETAP rating scale and the data itself constituted sources for creating certain categories out of the data (Kuckartz, 2014). In the first cycle of the data analysis, the subcoding strategy was employed to form subcategories within the research cases for main coding categories (Miles, Huberman, & Saldana, 2014). In the second cycle, pattern coding was applied to put the categories into themes with smaller numbers (Miles et al., 2014). Pattern coding allowed these categories to be inferred and explained across cases by constructing meaningful units. Each case was described narratively in detail and common patterns showing evidence for raters' rating behaviors were generated. After coding the data which was mostly in Turkish, the related parts were translated into English to be used as evidence in the result part.

3.7 Conclusion

In this chapter, the purpose of the research, data collection procedures, participant information, research questions and data analysis procedures are documented. In the next chapter, the results of the study are shared by reporting the analyses that are practiced to answer the research questions.

CHAPTER 4

RESULTS

The present study aimed to investigate whether ETAP raters showed any rater effects while assessing oral performances of test takers and to find out what their focuses were during the rating process. Five research questions were specified to obtain indications for rater severity/leniency, rater consistency, central tendency effect, rater bias towards criteria in ETAP scale and raters' foci in assessment process. In this chapter, analyses of the statistical data from Facets and the themes created with the help of NVivo 12 are reported for each research question separately.

4.1 Research question 1: Do ETAP raters differ in their scoring in terms of their individual level of severity and leniency when assessing ETAP test takers' oral performances?

In order to answer the first question, the Wright map (Wilson & Draney, 2000) has been checked and separation and reliability statistics of raters' measurement report are presented.

4.1.1 The Wright map

In the Wright map (Wilson & Draney, 2000), values for raters, test takers, criteria and scale are listed in logits (see Fig. 4). The most severe rater appears at the top of the column whereas the most lenient one is at the bottom. From the map, it seems that the most severe rater is Olcay while the most lenient one is Ege who shows the severity degree of around -3 logit. Deniz and Bahar appear in the middle of Olcay and Ege in the map and their severity levels are close to each other in logits in terms

of rater severity. Therefore, the severity degrees of Olcay, Deniz, Bahar and Ege decrease respectively.

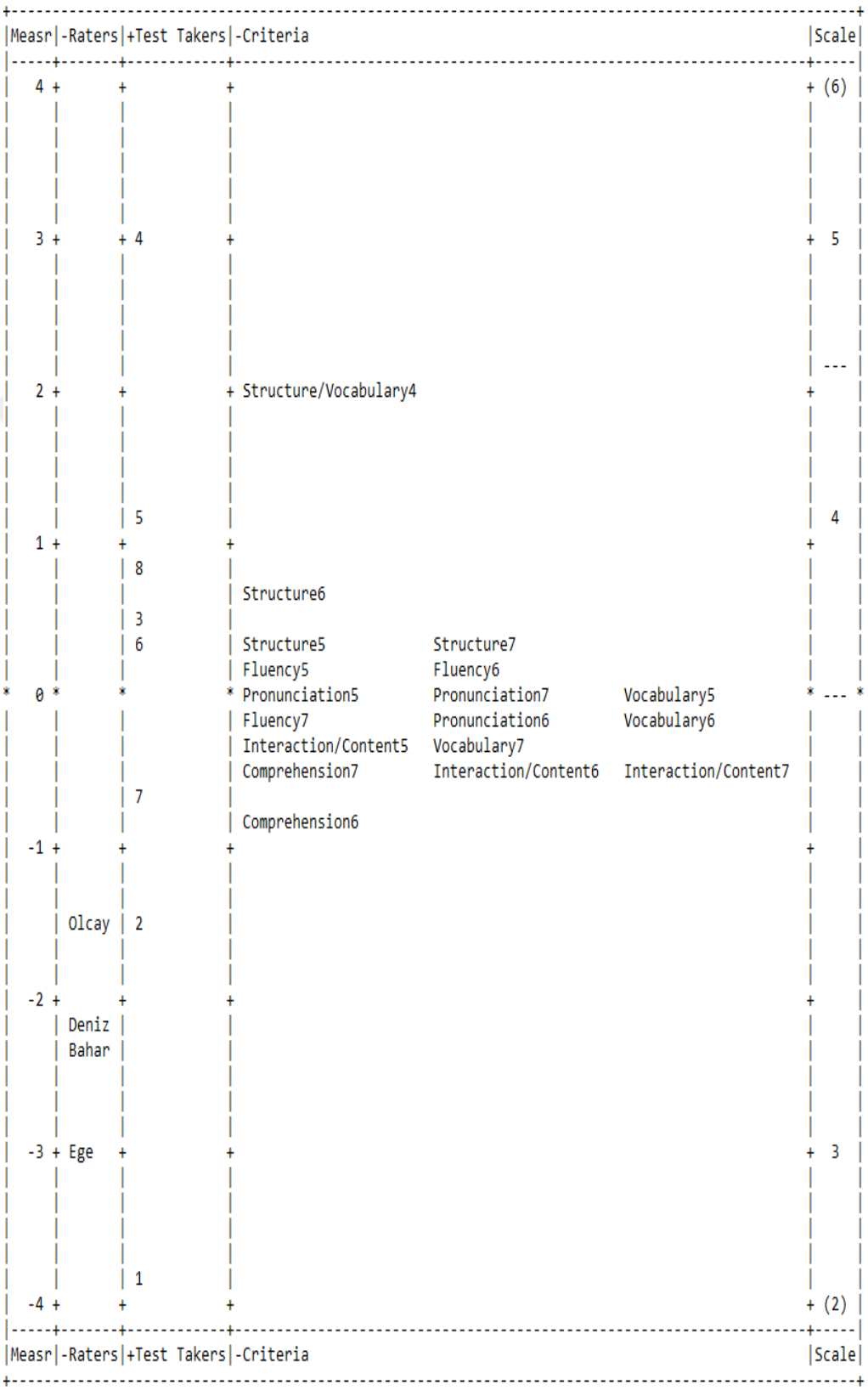


Fig. 4 ETAP raters' Wright map

4.1.2 Separation and reliability statistics

As in the Wright map, raters' measurement report (see Table 3) lists the raters from the most severe one to the least. At the right bottom of the table, the separation statistic is presented as 4.17 which displays that ETAP raters differ in severity. That is to say, there are at least four different severity levels among ETAP raters in assessing test takers' oral performances. Besides, the reliability level is 0.95 which, in turn, reveals that the raters differ significantly in their severity levels since the reliability statistic is close to the perfect separation value of 1. Therefore, it should be noted that ETAP raters are not interchangeable due to the significant difference in their severity levels.

Table 3. Raters' Measurement Report

Raters	Measure	Model SE	Infit MnSq	Outfit MnSq
Olcay	-1.47	.13	.94	1.25
Deniz	-2.17	.13	.87	.79
Bahar	-2.25	.13	1.02	.87
Ege	-3.04	.13	1.09	.97
Separation 4.17 Strata 5.90 Reliability (not inter-rater) .95				

4.2 Research question 2: How consistently do ETAP raters rate ETAP test takers' oral performances?

Although ETAP raters are different in severity, how consistent they are is important in terms of rater reliability. In order to understand whether the raters exhibit self-consistency in their ratings, infit and outfit MnSq values are investigated.

The infit and outfit MnSq values between 0.5 and 1.5 demonstrate raters' reliability in terms of their self-consistency in ratings (Linacre, 2002). The values under 0.5 indicate that raters are overpredicted so they use certain categories while evaluating performances and the values above 1.5 show that raters are misfitting so they are unpredicted in their ratings. In raters' measurement report (see Table 3), the

last two columns illustrate statistics for infit and outfit MnSq. The values for ETAP raters are between the reliable ranges of 0.5-1.5 which, in turn, means that all the raters show self-consistency in their ratings. That is to say, they are reliable in assessing test takers' oral performances.

The ZStd values are called as z-standardized and should fall between -2 and + 2 values. They can be easily inflated or deflated due to the small sample size and they are ignored when infit and outfit MnSq statistics are in the confidential range (Boone et al., 2014). Since fit statistics are between acceptable ranges, the ZStd values are neglected for this research question.

4.3 Research question 3: Do ETAP raters show any central tendency effect due to individual indistinguishability of score levels in ETAP rating scale?

Raters may overuse middle categories and avoid using extreme ones in rating scales (Myford & Wolfe, 2004). Rater infit and outfit MnSq values, partial credit models for raters and criteria, and test takers' separation statistics were analyzed one by one to determine whether ETAP raters demonstrate central tendency effect or not.

4.3.1 Rater Infit and outfit MnSq statistics

As an indication of central tendency effect, any value under 0.5 in rater infit and outfit MnSq statistics should be detected since it remarks that the raters are keen on using the middle categories in the rating scale and they are too predictable in their rating behaviors due to their overfitting values (Knoch, 2007). As stated in the results for the second research question, all ETAP raters are consistent in their ratings by locating themselves between acceptable ranges in terms of rater infit and outfit MnSq statistics. To make it more concrete, raters' measurement report (see Table 3)

displays that the raters Olcay, Deniz, Bahar and Ege own .94, .87, 1.02 and 1.09 infit MnSq values and they possess 1.25, .79, .87 and .97 outfit MnSq values respectively. That is to say, the rater infit and outfit MnSq statistics do not imply any central tendency effect for ETAP raters in general. However, for a deeper investigation for a central tendency effect, partial credit models for both raters and criteria were examined.

4.3.2 The first partial credit model '?,?,#'

This model demonstrates the usage of each criteria in detail. In this part, each criterion is examined based on their percentage in usage.

Firstly, Table 4 for the criterion structure illustrates the usage percentages for every category. In the descriptors for "Structure/Vocabulary4", it seems that the most used category is three with 44%. Later the category four and five come with 28% and 19% usage. The least used categories are two and six with 3% and 6% successively. For the criteria structure and vocabulary in Task 4, the categories two, three and six are extreme categories. According to the statistics, the raters tend to use the extreme category of three but not the categories of 2 and 6. In terms of structure in Task 5, the usage percentages are 22%, 34%, 19% and 25% for the categories of three, four, five and six respectively. Although the most used category is four, it does not show a big difference in usage when the usage percentages of other categories are taken into consideration. Therefore, it can be said that there is no central tendency effect for the criterion structure in Task 5. As for the criterion structure in Task 6, there is an even distribution in terms of category usage percentages. The category four is used most with 31% but the second and third most used categories are the extreme ones, three and six. Therefore, it means that there is no central tendency effect in terms of the

criterion structure in the Task 6. For structure in Task 7, Table 4 reveals similar results to the previous criterion. It shows an almost even distribution among the categories by exhibiting 25%, 31%, 16% and 28% for the categories three, four, five and six successively. This, in turn, is an evidence for the absence of central tendency effect.

Table 4. Partial Credit Model Analysis for Criterion Structure in ETAP

Categories	Counts Used	Percentage
Structure/Vocabulary4		
2	1	3%
3	14	44%
4	9	28%
5	6	19%
6	2	6%
Structure5		
3	7	22%
4	11	34%
5	6	19%
6	8	25%
Structure6		
3	9	28%
4	10	31%
5	6	19%
6	7	22%
Structure7		
3	8	25%
4	10	31%
5	5	16%
6	9	28%

In terms of the criterion vocabulary in Tasks 5, 6 and 7, Table 5 is presented. Statistics show that there is not a huge difference in the usage percentages for the categories of three, four, five and six so no central tendency effect for the criterion vocabulary in Tasks 5, 6 and 7 is shown.

Table 5. Partial Credit Model Analysis for Criterion Vocabulary in ETAP

Categories	Counts Used	Percentage
Vocabulary5		
3	5	16%
4	10	31%
5	9	28%
6	8	25%
Vocabulary6		
3	6	19%
4	7	22%
5	10	31%
6	9	28%
Vocabulary7		
3	4	13%
4	9	28%
5	11	34%
6	8	25%

As for pronunciation in Task 5 (see Table 6), the categories four and five are used eleven times while the categories three and six are utilized four and six times consecutively. The usage percentage for the categories four and five is 34% which, in turn, indicates the tendency towards these categories. Thus, there can be a central tendency effect for the criterion pronunciation in Task 5. However, the gap for category usage is not that vast in the Tasks 6 and 7 in terms of pronunciation even though the most utilized categories are four and five. Overall, the least used category for pronunciation is three whereas the most commonly used ones are four and five in Tasks 5, 6 and 7.

Table 6. Partial Credit Model Analysis for Criterion Pronunciation in ETAP

Categories	Counts Used	Percentage
Pronounciation5		
3	4	13%
4	11	34%
5	11	34%
6	6	19%
Pronunciation6		
3	4	13%
4	10	31%
5	10	31%
6	8	25%
Pronunciation7		
3	5	16%
4	10	31%
5	10	31%
6	7	22%

When it comes to fluency in Tasks 5, 6 and 7, the most commonly utilized category is five (see Table 7) in all tasks but the distribution among the categories is almost even so there is no clear evidence for central tendency effect in terms of fluency.

Table 7. Partial Credit Model Analysis for Criterion Fluency in ETAP

Categories	Counts Used	Percentage
Fluency5		
3	6	19%
4	9	28%
5	10	32%
6	7	22%
Fluency6		
3	8	25%
4	5	16%
5	12	38%
6	7	22%
Fluency7		
3	6	19%
4	7	22%
5	10	31%
6	9	28%

Table 8 for criterion comprehension in Tasks 6 and 7 exhibits a distinct classification in category usage. The usage of categories increases with their values. That is to say,

the least used category is three and then four. The frequently utilized ones are firstly six and then five. In Task 6, the usage percentage of the categories five and six is the same as 34%. Therefore, the statistical results report that ETAP raters are not keen on using the middle categories predominantly and escaping the extreme categories so no evidence of central tendency effect for comprehension is found.

Table 8. Partial Credit Model Analysis for Criterion Comprehension in ETAP

Categories	Counts Used	Percentage
Comprehension6		
3	3	9%
4	7	22%
5	11	34%
6	11	34%
Comprehension7		
3	5	16%
4	7	22%
5	9	28%
6	11	34%

Lastly, for the criterion interaction and content in Tasks 5, 6 and 7, the statistical results (see Table 9) support that the least utilized category is three in each task. In Task 5, the utmost utilized category is five with 41% but there is not a big difference in usage among the categories four and six. However, in Task 6 and 7, the categories five and six are used predominantly. Hence, it seems from the results that ETAP raters benefitted the extreme categories in rating interaction and content abilities of test takers.

Table 9. Partial Credit Model Analysis for Criterion Interaction/Content in ETAP

Categories	Counts Used	Percentage
Interaction/Content5		
3	4	13%
4	8	25%
5	13	41%
6	7	22%
Interaction/Content6		
3	4	13%
4	8	25%
5	10	31%
6	10	31%
Interaction/Content7		
3	5	16%
4	6	19%
5	11	34%
6	10	31%

4.3.3 The second partial credit model '#,?,#'

In the previous part, the central tendency effect is investigated with regard to each criterion thanks to the first partial credit model. Likewise, the second partial credit model reveals the usage percentages of each rater for each criterion. Thus, only the raters who display evidence for central tendency effect are represented in this part due to the large number of statistical results.

From the statistical results in Facets, the raters Deniz and Olcay show some evidence for central tendency effect in the criteria structure, vocabulary, pronunciation, comprehension, interaction and content. No evidence of central tendency effect for raters in terms of fluency is found in the results, which is parallel with the results attained from the first partial credit model.

Deniz displays central tendency effects for structure in Tasks 5, 6 and 7 (see Table 10). In Task 5, s/he uses the category four five times with 63% whereas s/he utilizes the categories three and six once and twice in sequence. S/he does not use the category five but it is clear that s/he has a tendency towards using the category four

in rating. In Task 6, s/he again shows the tendency by using the category 4 with 50% while the usage of the categories three and six is 25%. In Task 7, once again, s/he displays the tendency in rating towards using category four with 63% and the rest of the percentages for using categories three, five and six are the same in Task 5. S/he always avoids utilizing category five while evaluating structure so the reason for avoidance should be examined in detail with qualitative analyses.

Table 10. Deniz's Statistical Results for Criterion Structure

Categories	Task 5 (Used)	Task 5 (Percentage)	Task 6 (Used)	Task 6 (Percentage)	Task 7 (Used)	Task 7 (Percentage)
3	1	13%	2	25%	1	13%
4	5	63%	4	50%	5	63%
5	0	00%	0	00%	0	00%
6	2	25%	2	25%	2	25%

Olcay exhibits some evidence for central tendency so the statistical results for vocabulary, pronunciation, comprehension, interaction and content (see Table 11) are investigated. As for vocabulary in Task 7, s/he always uses the categories four and five while s/he does not use the extreme categories of three and six at all. This is clearly an indication of central tendency for vocabulary evaluation in Task 7. In terms of pronunciation in Task 5, 6 and 7, s/he uses the extreme categories three and six mostly just once and in Task 5, s/he abstains from utilizing the category six. However, the scoring is cumulated in the categories of four and five in three tasks in terms of pronunciation. Regarding comprehension criterion in Tasks 6 and 7, the scoring is gathered in the categories four and five by showing 38% usage for each whilst the usage for the extreme categories three and six is 13% separately. Finally for the interaction and content in Tasks 5 and 6, Olcay uses particularly the middle categories four and five. In Task 5, the uses of the categories three, four and five are 13%, 38% and 50% successively and there is no use of extreme category six, which

demonstrates the accumulation of middle categories in usage. In Task 6 for interaction and content, Olcay again avoids using the extreme categories and tends to use the middle categories of four and five with 38% for each.

Table 11. Olcay's Statistical Results for Central Tendency Effect

Criteria	Tasks	3 Used %	4 Used %	5 Used %	6 Used %
Vocabulary	7	0 00%	4 50%	4 50%	0 00%
Pronunciation	5	1 13%	4 50%	3 38%	0 00%
	6	1 13%	4 50%	2 25%	1 13%
	7	1 13%	4 50%	2 25%	1 13%
Comprehension	6	1 13%	3 38%	3 38%	1 13%
	7	1 13%	3 38%	3 38%	1 13%
Interaction/ Content	5	1 13%	3 38%	4 50%	0 00%
	6	1 13%	3 38%	3 38%	1 13%

4.3.4 Test takers' separation statistic

Test takers' separation statistics reveals their distinguishability in scoring. A high value is evidence for absence of central tendency effect since it signalizes the significant differences in their scores without any accumulation towards the middle categories. The table of test takers' measurement report (see Table 12) shows the separation value as 9.21 at the bottom, which means that there are more than nine different levels in eight test takers' measurements. That is to say, they are separated significantly in terms of ability. Additionally, this affirms the absence of central tendency effect as well.

Table 12. Test Takers' Measurement Report

Test Takers	Measure	Model SE	Infit MnSq	Outfit MnSq
4	3.06	.28	.63	.52
5	1.11	.17	1.15	1.16
8	.88	.17	1.26	1.21
3	.49	.16	.64	.65
6	.33	.16	1.46	1.50
7	-.67	.17	.83	.85
2	-1.46	.18	.65	.72
1	-3.75	.29	.91	1.16
Separation 9.21 Strata 12.61 Reliability .99				

4.4 Research question 4: Do ETAP raters show any rater bias towards any criteria in the rating scale? If yes, for which criteria do the raters behave more severely or leniently?

'Rater x Criteria' bias/interaction analysis was carried out to learn whether ETAP raters display any bias against any criteria in ETAP rating scale. The statistical analyses demonstrate the bias size in t-values. The highest bias rate is shown on the top of the table. The difference between observed score and expected score show the bias size converted in logits in the table.

The bias/interaction analyses for ETAP raters reveal that there is no rater showing bias towards any criteria in the rating scale since all t-values are between the acceptable ranges of +2 and -2 (McNamara, 1996; Lumley & McNamara, 1995; Knoch et al., 2007). However, the severity rates for both interrater and intra-rater assessments change when the pairwise reports are checked. The first bias/interaction pairwise report (see Table 13) reveals that Deniz rates the criterion structure and vocabulary in Task 4 more leniently than Olcay and Ege and the differences between ratings are significant (t = 2.50 and t = 2.49). Likewise, Bahar evaluates the same criterion more leniently than Ege and Olcay (t = -2.16 and t = -2.18).

Table 13. The First Bias/Interaction Pairwise Report

Criteria	Obs-Exp Average	Rater	Obs-Exp Average	Rater	t-value
Structure/Vocabulary4	-.41	Deniz	.33	Olcay	2.50
Structure/Vocabulary4	-.41	Deniz	.38	Ege	2.49
Structure/Vocabulary4	.38	Ege	-.31	Bahar	-2.16
Structure/Vocabulary4	.33	Olcay	-.31	Bahar	-2.18

As for the second bias/interaction pairwise statistics (see Table 14), Deniz rates the criterion interaction and content in Task 7 more severely than structure and vocabulary in Task 4 (t = 2.01). On the contrary, Ege assesses interaction and content in Task 7 more leniently than structure and vocabulary in Task 4 (t = -2.02).

Table 14. The Second Bias/Interaction Pairwise Report

Rater	Obs-Exp Average	Criteria	Obs-Exp Average	Criteria	t-value
Deniz	-.41	Structure/ Vocabulary4	.21	Interaction/ Content7	2.01
Ege	.38	Structure/ Vocabulary4	-.25	Interaction/ Content7	-2.02

4.5 Research question 5: What do ETAP raters focus while rating oral performances?

The verbal reports of raters contributed to answer the fifth question regarding the focuses of raters while rating the examinees. More than 20 patterns were created out of the data. However, the issues that at least three raters discussed are provided in the results since they are the most common comments of the raters. Therefore, the most frequent focuses of the raters are:

1. Accent and sharing the same mother tongue with examinees
2. Testing process and the interlocutor in the exam
3. The rating scale
    a. Criticism about the rating scale

b. Rating scale alteration

c. Referring to the rating scale while rating

4. Criteria

    a. Criterion related comments

    b. Criterion unrelated comments

    c. Impacts of some criteria on others

5. Appropriateness and meaningfulness of answers

6. Comparison with other test takers and across tasks

Each of the patterns is explained and excerpts from the data are provided below.

4.5.1 Accent and sharing the same mother tongue with examinees

ETAP raters mentioned accents of Turkish pilots in their verbal reports. They highly commented on the ways the pilots pronounced specific words and sounds. Deniz, Ege, Olcay and Bahar comment on six, four, two and five examinees' accents successively. Besides, raters considered the intelligibility of the pronunciation and sharing the same L1 with the examinees while rating the performances. Except for Ege, all raters mention having the same mother tongue with the test takers. However, they refer to it only once for three different examinees.

> Deniz: When we look at the Task 5 and start from the pronunciation, the Turkish accent is felt extremely especially for 'th' and 'w' sounds, rhythm and intonation. Some pronunciations can cause problems like 'noses' and 'collide with'. I could not hear the word 'bird' clearly. [The pilot] has a different pronunciation throughout the recording but generally [this] does not cause any serious ambiguity. So, I think [the examinee] has reached Level 4 in terms of pronunciation. (Excerpt 1)

> Ege: The only problem was [his] pronunciation. His pronunciation was a little bit bad but it does not hinder intelligibility. I do not think he can correct it [his pronunciation] since there is a concept of World Englishes. (Excerpt 2)

> Olcay: For this task, I assigned three [points] for pronunciation because he [the examinee] pronounces all the words, which appear in Turkish as well, like in Turkish without using [British/American] accent so this [can] create hindrance against intelligibility for any English speakers except for us [Turkish speakers of English]. So, I gave three [points]. (Excerpt 3)

> Bahar: I can understand [his accent] because I am Turkish but a foreign pilot or ATC might not understand him. (Excerpt 4)

In general, the raters do not see the accent differences as fatal mistakes in pronunciation as long as the examinees' accents are understandable by speakers of English. Ege states that s/he is aware of the concept of world Englishes, the focus of which is English varieties (Kachru, 2003), so s/he accepts the pronunciation of the examinee as an operational level. Moreover, Olcay and Bahar focus on the issue of sharing the same mother tongue with the examinees. They think that they can understand the utterances of the examinees since they all share the same L1. Therefore, they try to look from the point of other English speakers as well. By taking this into consideration, the raters grade the oral performances of the pilots.

4.5.2 Testing process and the interlocutor in the exam

ETAP raters have different ideas regarding the testing process but they share similar comments for the interlocutor in the exam. Deniz, Ege and Bahar share their ideas about the interlocutor and the testing process of six examinees. Firstly, the ideas for the testing process are shared and then the comments for the interlocutor are provided.

> Deniz: [The examinee] misunderstood one question. Generally, it is misunderstood… when [the examinee] misunderstands the question, the interlocutor asks the question again and [the examinee] can answer the question correctly. I have some concerns at this point. To what extent is it appropriate to explain the question [again] as for giving us idea [regarding the proficiency of the examinee]? Of course, there is something wrong with the question… Perhaps, the formation of the question can be changed… After the interlocutor reforms the question, [the examinee] answers it correctly. Therefore, I did not take points off here and graded as five. (Excerpt 5)

> Ege: The descriptions for the question were not provided completely. The fact that [you] should ask three questions was not delivered directly. This had an impact on the candidate [the examinee]. (Excerpt 6)
>
> Bahar: In Task 5, [the examinee] did a very long explanation. So, I think there should be a time limitation for answers. (Excerpt 7)

For the testing process, the raters think that instructions should be delivered clearly. Whether the interlocutor explains the questions twice or reformulation of some questions is needed should be decided before the examination. Besides, putting a time limit for answers is another suggestion from the raters.

ETAP raters state the mistakes of the interlocutor, who conducts the test, as well.

> Deniz: In Task 7, the interlocutor has serious problems in terms of pronunciation and this affects the examinee's pronunciation, too… I think the performance of the interlocutor has a negative influence on the [examinee's] pronunciation so I gave four points here. (Excerpt 8)
>
> Ege: There are some little mistakes of the examiner, too. (Excerpt 9)
>
> Bahar: [The test taker] mispronounces 'procedure' in Task 7. Of course, this can be resulted from the examiner who mispronounces it as far as I remember. However, the examiner should be more careful about that. (Excerpt 10)

The raters think that the reasons why the test takers mispronounce certain words can be due to the pronunciation mistakes of the examiner. Therefore, the interlocutors should be more careful while conducting the test since the scores can be influenced by this variable.

4.5.3 The rating scale

While rating, raters evaluate the performances that tasks require by depending on the rating scale so they need to find the parallelism between the performances and the criteria defined in the scales. When raters have difficulty in detecting the analogy, they can criticize the rating scales and demand certain alterations. Except for Bahar,

ETAP raters report their critique regarding the rubric. Deniz and Ege talk about the rating scale twice while assessing different examinees, Olcay mentions it three times.

> Deniz: I think [Task 4] is difficult to assess because vocabulary and structure are supposed to be evaluated together. Since we cannot evaluate structure independently, this can create problems in scoring. (Excerpt 11)
>
> Ege: There are some questions that he did not understand in Task 4 but we do not assess comprehension in Task 4. (Excerpt 12)
>
> Olcay: I gave four points in Task 4 because I know that we do not assess [the performance] in terms of comprehension but most of the questions that [the examinee] asked were a little bit nonsense. (Excerpt 13)

The common problem stated by the raters was about the criteria for Task 4. The joint assessment of structure and vocabulary can cause difficulties in rating since the vocabulary usage or the control on structure can be better than the other. In this case, it is hard for raters to assign a fair score. Moreover, raters think that comprehension can be assessed in Task 4 since examinees should understand the cases so that they can ask appropriate questions to get details about the incident.

ETAP raters do not directly state the alteration of the rating scale but they have certain comments indicating a need for creating a new category in the middle of two descriptors of a criterion. Deniz reports it three times in her verbal recordings and Olcay imply it once.

> Deniz: Here, [the examinee] is better in terms of grammar when compared to other tasks. Actually, we can consider it as four plus (+) points but I do not think that he gets five [points] because there are mistakes in tenses. (Excerpt 14)
>
> Olcay: In Task 4, I gave four points to vocabulary and structure because the questions he [the examinee] asks are not exactly for five points but a little less. (Excerpt 15)

From the excerpts, it seems that raters are doubtful about the score that they assign. They generally avoid assigning the higher point when they are indecisive between two scores. They think that giving the higher grade is not appropriate for the shown performance but giving the lower score might not be fair as well. That is why they

state a plus (+) score between the higher and the lower points. In that case, the descriptors in the rating scale can be revised or another category can be added into the rating scale.

Rating scales are guidance of raters so they often refer to the descriptors in order to decide on the scores they assign. While Deniz, Olcay and Bahar report it four times in their different assessment processes, Ege refers to the rating scale only once in the verbal recordings.

> Deniz: The errors in the structure are getting attention. I cannot decide on whether to assign three or four points. However, when we check the descriptor, the important issue is to what extent the errors interfere with the meaning. When I evaluated [structure] according to the descriptor in the scale, I assigned the operational score four to the structure. (Excerpt 16)

> Olcay: I gave the point five for the structure but I hesitated to give four or five. However, after reading the scale, for the Level 5, it says 'complex structures are attended but with errors.' So, it is not four… but at least [the examinee] can attend complex sentences with some errors so I think that assigning five points is fairer. (Excerpt 17)

> Bahar: I think errors do not interfere with the meaning. [In the rating scale] it says 'speaker can use complex structures only with negligible errors.' Due to the stated 'negligible errors', I gave five [points]. (Excerpt 18)

As stated in the excerpts above, ETAP raters stick to the rating scale while evaluating the performances. They try to find clues in the descriptors to rate the examinees when they hesitate the score they assign.

4.5.4 Criteria

The criteria and the descriptors in the ETAP rating scale were taken into consideration while analyzing the criterion related and unrelated comments in the verbal reports. Therefore, the raters' evaluations regarding pronunciation, structure, vocabulary, fluency, comprehension and interaction were coded separately. The relatedness of the comments was determined by depending on the explanations in the descriptors (see in Appendix A). Table 15 reveals the details of both criterion

relevant and irrelevant comments. Furthermore, the irrelevant ones are explained in detail by providing sample excerpts.

Table 15. Raters' Criterion Relevant and Irrelevant Comments

Criteria (Total)	Criterion related comments (References/Times)	Criterion unrelated comments (References/Times)
Pronunciation (80)	Intelligibility (28) L1 influence (16) Pronunciation mistakes (27) Intonation (5)	Murmuring (4)
Structure (165)	Articles (4) Complex structure (12) Control on structure (17) Gerund & Infinitive (1) Grammar mistakes (68) Intelligibility (19) Parts of speech (7) Passive (5) Plural (3) Relative clauses (1) Reported speech (1) Tense (20)	Hesitations/fluency (4) L1 influence (2) Low frequency in usage (1)
Vocabulary (75)	Chunks (5) Control on vocabulary usage (1) Idioms (1) Misusage (13) Phrasal verbs (4) Sufficient vocabulary (10) Terminology (3) Low frequency words usage (7) Vocabulary range (30)	Pronunciation mistakes (1)
Fluency (67)	Fillers (14) Hesitations (22) Intelligibility (10) Natural speaking flow (5) Speaking speed (14) Wrong use of connectors (1)	Slip of tongue (1)
Comprehension (42)	Answers to the questions (5) Summary (18) Understanding the questions (16)	Clarification questions (3)
Interaction (40)	Clarification questions (2) Detailed answers (16) Fluent-smooth interaction (6) Summary (8)	Hesitations (1) Understanding questions (7)

ETAP raters are consistent in rating, which means that they have a shared understanding in terms of the construct to be assessed and they follow the rating scale properly while rating test takers' performances. They make interpretations about the performances by mostly benefiting from their own criterion relevant comments (see Table 15). Firstly, raters comment on criterion pronunciation 80 times in total and the percentage of following the descriptors in the rating scale is 95. Secondly, since Task 4 only demands the correct usage of structure, it seems that comments on it outnumber the other comments for the other criteria. Raters' references to the structure consist of almost 96% relevant comments. Thirdly, raters refer to the criteria vocabulary and fluency and their comments are approximately 99% relevant to the ETAP rating scale. Lastly, while raters verbalize 93% relevant comments regarding criterion comprehension, their references to criterion interaction are 80% related to the descriptors in the rating scale. In total, raters refer to the criteria 469 times and their 24 references are irrelevant to the specific descriptors in the rating scale. Therefore, raters utter approximately 95% criterion relevant comments in their verbal reports. That is to say, the findings of the qualitative analyses are in line with the quantitative analysis indicating that ETAP raters follow mainly the descriptors in the ETAP rating scale while assessing performances so their rating behaviors are consistent.

From the Table 15, it seems that ETAP raters mostly follow the descriptors in the ETAP scale while rating examinees. However, there are some criterion unrelated comments made by the raters while assessing certain skills. For the criterion pronunciation, there is no explanation regarding murmuring in the ETAP descriptors however the raters Ege and Bahar state this in their oral reports respectively once and twice.

> Ege: In pronunciation, I did not count [his] murmurs and I thought that this was his way of speaking so I gave five points. (Excerpt 19)
>
> Bahar: Unfortunately, the pronunciation is three because he [the examinee] is very silent and he is murmuring words. (Excerpt 20)

As stated in the excerpts, murmuring did not affect the examinee's score for Ege but Bahar reports that there was an effect of murmurs on scores.

Hesitations/fluency, L1 influence and low frequency in usage are other criterion unrelated comments in structure. While Deniz and Olcay comment on hesitations and smoothness once, Ege refers to it twice for different examinees' structure. Olcay and Bahar state the influence of the mother tongue on structure only once for different examinees. Furthermore, Deniz talks about the low frequency in structural usage once in the recordings. For each, an excerpt is shared successively.

> Olcay: Although he [the examinee] utilizes a lot of structures incorrectly, he speaks so fluently and confidently that what he says is understood. Actually, the way he speaks compensates his mistakes in structure so it is not fair to take points off structure. So, I did not take points off and gave five points. (Excerpt 21)
>
> Bahar: When we look at Tasks 6 and 7, [the examinee] uses L1 structure. There is an L1 influence on both vocabulary and pronunciation but it also affects structure. Here, he says something like 'enough time before'. This shows that L1 structure is effective in his L2. (Excerpt 22)
>
> Deniz: 'Did you clear the area around the snake's position'. Here, the usage of 'area around the snake's position' is not common [in structure]. So we can say that there is failure here. (Excerpt 23)

As seen in the excerpts, the way examinees speak, their fluency and hesitations in their speaking can impact the assessment of structure. Raters also pay attention to examinees' errors and they deduce the reasons of the errors. For instance, Bahar reckons that the structural errors of the examinee are due to L1 influence. Moreover, while rating the performances, raters also behold the common structural usages in the target language. If a grammar structure is used rarely, it can be counted as errors.

Considering pronunciation mistakes is not relevant to vocabulary criterion in the ETAP rating. There is only one comment about it and it affects the examinee's vocabulary score.

> Deniz: The word 'upgrade' is a good vocabulary choice but [the examinee] mispronounced it. (Excerpt 24)

Slip of tongue is not expressed in the ETAP descriptors for fluency but Deniz is the only rater who mentions it in the verbal reports. However, s/he also states other reasons for taking points off.

> Deniz: We come across hesitations, slip of tongue and wrong beginnings a lot. Fillers and the slowness in speaking speed decrease the functionality of fluency. So, I graded [fluency] three and below. (Excerpt 25)

In the ETAP rating scale, clarification questions are evaluated for the criterion interaction but one rater uses them to assess comprehension. She mentions clarification questions three times for different tasks of the same examinee and she combines the assessment of comprehension and interaction.

> Bahar: As opposed to the questions of the interlocutor, he [the examinee] asks what she has meant or where she wants him to start. This shows me his high [proficiency] in comprehension and interaction. (Excerpt 26)

As for interaction, Ege considers that hesitations in speaking affects the candidate's interaction score. This appears only once in the recordings. Moreover, Deniz, Olcay and Bahar report that understanding the questions affects the score of interaction while assessing oral performances.

> Ege: When we look at interaction in general, I want to give five points because hesitations affect the fluency in interaction. (Excerpt 27)

> Olcay: Towards the end, he [the examinee] had serious problems in interaction. He did not understand two follow up questions. (Excerpt 28)

Even though the unrelated criterion comments are not frequent among raters, they need to be paid attention and raters should be informed regarding their judgements so that each criterion can be assessed validly and consistently.

In addition to the criterion related and unrelated comments, raters mention the impacts of some criteria on others. Deniz notifies the impacts twice for different examinees' performances and the rest of the raters mention it once.

> Deniz: The limited vocabulary range and mistakes in structure cause tension and anxiety in [his] speaking. We can feel that. These tension and anxiety are important factors for hesitations so we come across a bunch of hesitations during the recording. Especially, there are unfinished words, wrong beginnings and pauses throughout the recording. This affects fluency badly and I graded it three and below because I think that [the examinee's performance] is not operational. (Excerpt 29)

> Ege: If his grammar were better, I could give five points to his vocabulary but his deficiency in grammar affects his vocabulary as well so I wanted to give four points for vocabulary. (Excerpt 30)

> Olcay: His slow pace in speaking causes grammar mistakes as well so I gave three points. (Excerpt 31)

> Bahar: Because his structure is not automatic, he [the examinee] thinks a lot and hesitates. (Excerpt 32)

In general, ETAP raters deem that the lack of performance in vocabulary and structure results in failure in terms of performances in fluency, comprehension and interaction. They also reckon that structural production affects vocabulary usage, and fluency in speaking has impact on structure as well. To ascertain the impacts of criteria on each other, the examinees' performances should also be analyzed linguistically.

4.5.5 Appropriateness and meaningfulness of answers

In addition to six criteria in ETAP rating scale, raters evaluate both appropriateness and meaningfulness of examinees' answers. In Task 4, test takers should ask fifteen different questions related to the given situations and in other tasks, they are supposed to answer the follow up questions properly. Therefore, ETAP raters need to judge whether the given answers are convenient and fulfill the requirements in the

rating scale. In this respect, Deniz delivers her/his comments six times for different examinees whereas Ege, Olcay and Bahar discuss it three times.

> Deniz: In the fourth case, there is a problem about primary flight display. [The examinee] asks questions like ‘what is the problem exactly?’, ‘how much time do you have this problem for?’ and ‘is the other primary flight display working on that?’ The problem is interrogated firstly. There is no error in structure here… I do not understand his last question. What does he imply by saying ‘primary flight display’? I know that there are two [primary flight displays] in planes. He ends his question with ‘working on that’. I do not quite understand what is stated here. (Excerpt 33)

> Olcay: I think that the questions he [the examinee] asked were not appropriate and precise in Task 4. In my opinion, he could not ask the questions he should ask. (Excerpt 34)

> Bahar: In Task 4, a question like ‘what is the color of the snake?’ was formed as the first question. I think it was an unrelated question because we do not have anything to do with the color of the snake. (Excerpt 35)

As seen from the excerpts, raters comment on the convenience of the examinees’ answers. However, they sometimes are not sure of it due to their unfamiliarity with the aviation field. In order to judge the appropriateness and meaningfulness of the examinees’ answers, raters should be informed about the given incidents and the questions in the test beforehand. Additionally, they can be provided with a sample answer key so that they can have extensive understanding about the topic in the field.

4.5.6 Comparison with other test takers and across tasks

ETAP raters occasionally compare the examinees with the ones they have assessed before. It is found that all raters except for Ege make comparison of test takers’ performances. Deniz compares the examinees twice throughout the verbal reports while Olcay and Bahar do it three times.

> Deniz: Especially in this part, I really like his [examinee’s] comprehension and interaction performance. He was the only candidate who gave the exact reason of the incident in comparison with the other recordings [examinees] that we had listened to. He summarized the incident very well. He answered all the questions sufficiently and in detail. Therefore, I gave six points for comprehension and interaction in the last task, too. (Excerpt 36)

> Olcay: I think that it would be unfair to give five points for his [examinee's] structure because there were other test takers whose performances in structure were better so I think his structure is a good four but not five. (Excerpt 37)

> Bahar: His vocabulary usage is very good. I mean it is better than the captain's that I listened to previously. (Excerpt 38)

As narrated in the excerpts, raters rely on their judgements by occasionally comparing examinees' performances. They reckon that they compare the performances to assign fair scores. However, ETAP is a criterion-referenced test so performances are assessed by depending on the descriptors in the rating scale. A comparison among test takers risks the nature of the test and can turn it into a norm-referenced test where performances are interpreted with respect to other examinees' performances (Ang-Aw & Chuen Meng Goh, 2011). Therefore, comparisons between examinees should be abstained.

In addition to the examinee comparison, raters make comparison of the same test taker's performances in different tasks. Deniz, Ege, Olcay and Bahar compare the performances in different tasks respectively five times, twice, four times and once.

> Deniz: Here, he [the examinee] performs a little bit better than the other tasks in terms of grammar. (Excerpt 39)

> Ege: In Task 5, his [the examinee's] grammar usage was better than the one in Task 4. (Excerpt 40)

> Olcay: I gave four points to his [the examinee's] vocabulary because his vocabulary range was limited in Task 5 although his vocabulary range was wider in Task 4. (Excerpt 41)

ETAP raters try to spot abilities and failures of test takers. While practicing it, they realize the differences in performances of an examinee throughout the test. In ETAP, different tasks aim to evaluate certain criteria in test takers' performances. Therefore, the descriptors in the rating scales for different tasks are similar but there are tiny

differences to assess the targeted construct. Therefore, raters should refer to the descriptors while comparing the same criteria in different tasks.

4.6 Conclusion

This part provides both quantitative and qualitative analyses for the intended research. For the first four research questions, statistical results are delivered in tables from Facets and for the last question, a detailed qualitative analysis is presented with common themes and categories. A thorough discussion for the given results is held in the next chapter.

CHAPTER 5

DISCUSSION

A mixed methods approach including many-facet Rasch analysis for the quantitative part and NVivo12 for the qualitative part of the study is utilized for this research on the purpose of detecting rater effects and focuses in ETAP rating process. In the previous chapter, raters' severity levels, individual consistency in rating, central tendency and bias against criteria in the rating scales are documented at first and then, raters' focuses on rating ETAP test takers are reported in detail with sample excerpts from the raters' verbal reports. In this chapter, the results of the analyses are discussed for each research question.

5.1 Research question 1: Do ETAP raters differ in their scoring in terms of their individual level of severity and leniency when assessing ETAP test takers' oral performances?

One of the irrelevant variables to the constructs aimed to be tested is raters' severity in rating. Raters may exhibit severity or leniency effect while rating certain examinees, task and criteria (Knoch et al., 2007). In the first research question, raters' severity levels are investigated to find out whether raters are interchangeable in rating ETAP examinees. In order to answer this question, Wright map, separation and reliability statistics in the raters' measurement report were examined. It is found that there are at least four different severity levels that ETAP raters display. The severity levels of Ege, Bahar, Deniz and Olcay increase successively. The results mean that raters exhibit significantly different severity levels in rating test takers' performances so they are not interchangeable.

It seems that even after the online rating training, the severity degrees of ETAP raters vary. On this point, Weigle (1998) highlights that differences in severity levels can continue even after rater trainings. Therefore, ETAP raters' differences in severity levels can be acceptable as natural. However, the important question is whether the considerable differences in severity degrees of raters lead to any problems in ratings or not? According to Eckes (2009), this does not cause any difficulty when raters' rating behaviors are consistent throughout rating process. Therefore, more importance should be attributed to individual rater consistency than rater severity. In the second research question, rating consistency is taken into consideration and explained by referring to the severity degrees as well. Results show that ETAP raters are consistent in their rating behaviors so the difference in their severity levels do not create any problem for the rating process.

Moreover, there are ways to eliminate the raters' severity effects on scores. Eckes (2005) states that examinees' fair scores, which can be attained by means of MFRM, can constitute their final scores since the fair scores are the adjusted observed scores for the variation in raters' severity degrees. Apart from score adjustment with statistics, Myford and Wolfe (2003) suggest that raters should be aware of the impacts of the severity degrees on scores and they should be trained to specify categories in rating scales precisely in order to decrease severity effects on ratings. Furthermore, they recommend that more than one rater can evaluate the same examinee so that an average score can be gained from both severe and lenient raters.

5.2 Research question 2: How consistently do ETAP raters rate ETAP test takers' oral performances?

Another rater variable is inconsistency effect which implies that raters are not consistent in their rating behaviors (Knoch et al., 2007). That is to say, it demonstrates whether raters are reliable in rating or not. If raters are consistent in their rating behavior, they share the same understanding in terms of the construct aimed to be tested (Yan, 2014). In order to learn raters' self-consistency in rating, the values of rater infit and outfit statistics were investigated. Linacre (2002) affirms that if the fit statistics are within the acceptable ranges (between 0.5 and 1.5), the raters are consistent in rating so they are reliable. Likewise, the MFRM analyses illustrate that fit statistics are between acceptable ranges for each ETAP rater. That means all ETAP raters are consistent in their rating behaviors and they utilize the rating scale in similar ways so they are reliable in rating.

Weigle (1998) claims that rater training is useful for increasing rater consistency in rating, decreasing differences in rating severity levels and rater bias. Similarly, but with a tiny difference, McNamara (1996) underlines the acceptance of inter-rater variability and individual rater consistency in ratings. Therefore, intra-rater consistency should be the most appropriate purpose of rater trainings. Moreover, Eckes (2009) states that high level of inter-rater agreement (inter-rater reliability) on scores can be misleading since this does not guarantee accuracy of ratings. That is to say, rater consistency is a more convenient way to find out the proper application of rating scales by raters. As discussed in the first research question, the severity levels of ETAP raters differ but their individual consistencies exhibit that they apply ETAP rating scale similarly so they are reliable in rating. Furthermore, Davidson (1991) and North (2000) (as cited in Li & He, 2015) add that

raters' consistency in understanding and applying rating scales is significant for validation of a rating scale. That is to say, the consistency of ETAP raters in rating can be beneficial for the development and validation of ETAP rating scale for future studies.

5.3 Research question 3: Do ETAP raters show any central tendency effect due to individual indistinguishability of score levels in ETAP rating scale?

Raters may tend to overuse the middle categories in the rating scales and they can avoid using the extreme categories while rating performances. When they are consistent in this rating behavior, they may show central tendency effect (Myford and Wolfe, 2003). In order to detect the central tendency effect, raters' fit statistics, partial credit models for both raters and criteria, and separation statistics of examinees were investigated via Facets. Raters fit statistics are between acceptable ranges and imply that there is no central tendency effect among ETAP raters. In terms of criteria usage examined through a partial credit model, the absence of central tendency effect for the criteria structure, vocabulary, fluency, comprehension, interaction and content was found. However, results indicate that there can be a central tendency effect for the criterion pronunciation in Task 5. Furthermore, the criteria usage by each rater was analyzed with a second partial credit model in Facets. Results reveal that Deniz is keen on using the category four for the criterion structure in Tasks 5, 6 and 7. However, s/he does not utilize the category five whereas s/he uses the extreme categories three and six. Apart from the central tendency effect, no usage of category five by Deniz should be investigated thoroughly in future studies. Olcay displays central tendency effect for the criteria vocabulary, pronunciation, comprehension, interaction and content by using the

category four and five mostly. Eventually, examinees' separation statistics were reviewed to ascertain the distinguishability in scoring. Results reveal that examinees' abilities are separated significantly without any accumulation in the middle categories, which unveils the absence of central tendency effect in rating.

Although the fit statistics and examinees' separation statistics approve the absence of central tendency effect, the statistics in the partial credit models signalize the existence of central tendency effect on certain occasions. Which statistical result is more reliable in this case? The fit statistics showing overfit means that raters are too predictable in their ratings. The scores they assign can be estimated in general since they can be keen on using the middle categories at most (Knoch, 2007). However, Myford and Wolfe (2004) emphasize that fit statistics sometimes might not indicate clear results for the central tendency effect. To be more precise, partial credit models should be applied since they reveal more in-depth analyses by showing individual usage of each category in each criterion. Besides, a high value in test takers' separation statistics signal the absence of the central tendency effect (Sudweeks et al., 2005) but it does not provide detailed statistics regarding the utilization of each category. Therefore, as Myford and Wolfe (2004) indicate, partial credit models provide complete analyses for the inspection of the central tendency effect. However, so as to detect raters' persistence in the central tendency effect via statistical analyses, ETAP raters should rate more test takers and examinees' previous ICAO scores should be known to make comparisons.

Apart from raters' rating patterns, the central tendency effect can be an indication of a rating scale variable. For instance, in their article, Myford and Wolfe (2004) claim that a revision can be required for categories in a rating scale if the central tendency effect exists in many raters' ratings. Fortunately, ETAP raters

mostly do not display any central tendency effect and this implies no need for any modification in categories of ETAP rating scale.

Raters may display central tendency when they know that their ratings are monitored (Myford & Mislevy, 1995, as cited in Knoch et al., 2007) since they want to avoid making mistakes while rating. Therefore, it is probable that they show less central tendency effect in the circumstances when they are not observed. In a similar vein, Wolfe and McVay (2010) find out that raters trained online can show less central tendency effect in scoring. That is to say, online training of ETAP raters might contribute to the infrequent central tendency effect among raters but this requires further research including a control group which takes a face to face rater training.

All in all, in order to prevent central tendency effect, Myford and Wolfe (2003) recommend that there should be clearly described categories in rating scales for raters to be able to differentiate levels of performances. Besides, raters should be made aware of the central tendency effect and its impact on scores. In rating training sessions, raters can also be forced to use each category in a rating scale to rate predetermined performances of examinees and to rank them in order.

5.4 Research question 4: Do ETAP raters show any rater bias towards any criteria in the rating scale? If yes, for which criteria do the raters behave more severely or leniently?

Raters can exhibit bias towards certain criteria by rating them consistently more severely or leniently. In order to find out whether ETAP raters show any bias towards certain criteria in ETAP rating scale, ‘Rater x Criteria’ bias/interaction analysis was conducted. When the bias/interaction table for raters was checked, no

rater bias was found because all the t-values were between the acceptable ranges of +2 and -2 (McNamara, 1996; Lumley & McNamara, 1995; Knoch et al., 2007). However, pairwise comparison analyses demonstrate certain bias against particular criteria in terms of interrater and intra-rater assessments. As for interrater comparison, Olcay and Ege rate the criterion "Structure/Vocabulary" in Task 4 more severely than Deniz (t = 2.50, t = 2.49) and Bahar (t = 2.18, t = 2.16) and there are significant differences between the ratings. In terms of intra-rater assessment, Deniz rates the criterion "Structure/Vocabulary" in Task 4 more leniently than "Interaction/Content" in Task 7 whereas Ege rates the criterion "Structure/Vocabulary" in Task 4 more severely than "Interaction/Content" in Task 7.

Before conducting the analyses, the researcher of this thesis concerned that there was a bias against the criterion structure among ETAP raters. The Facet results show that she was partially right in her observations during the rating training. ETAP raters Olcay and Ege exhibit a bias against "Structure/Vocabulary" in Task 4 by rating it more severely. Moreover, the intra-rater comparison analyses demonstrate that Deniz and Ege apply significantly different severity levels while rating the criteria "Structure/Vocabulary" in Task 4 and "Interaction/Content" in Task 7. Likewise, in the MFRM analyses of OET, McNamara (1996) found out that test takers' grammatical performance affected raters' judgements and raters were not aware of the influence. In other words, raters displayed bias towards grammatical accuracy unconsciously although the test consisted of a communicative base. Similarly, Lumley (2005, as cited in Schaefer, 2008) ran MFRM analyses to investigate rater bias in the Special Test of English Proficiency (STEP). Four trained STEP raters rated 24 writing tasks selected from the pool of the test. It was found

that STEP raters was biased against the criterion grammar. Moreover, Eckes (2012, as cited in Shirazi, 2019)) conducted a 'Rater x Criterion' bias analysis for 18 ratings to investigate the connection between rater behavior and cognition. He found out that when raters considered certain criterion as quite significant, they rated it severely; otherwise, they rated it leniently. That is to say, Olcay and Ege might think that the criterion "Structure/Vocabulary" is significant while rating the performances in Task 4 since the actual focus of Task 4 is the accurate usage of grammar and vocabulary. Likewise, Task 7 aims to assess examinees' communicative competence so Deniz may consider that the criterion "Interaction/Content" is more crucial for that task.

Examinees' scores can be affected by bias scores assigned by raters, which decreases validity of a rating process. However, validity can be maintained through MFRM analysis since it yields accurate and reliable estimation of test takers' abilities by excluding rater effects on scores (Linacre, 1989). Apart from validity of a rating process, bias analyses are beneficial to detect raters' systematic patterns in rating so that they can contribute to development of rating scales and rater trainings (Schaefer, 2008). Wigglesworth (1993, as cited in Schaefer, 2008) examined the benefit of bias analyses in rater trainings. She provided raters with the MFRM bias reports in rater training sessions. Thanks to the feedback, it turned out that raters decreased their bias and improved their consistency in rating. In a similar vein, the MFRM bias analyses can be utilized for individual rater feedback in the future ETAP rater training sessions.

5.5 Research question 5: What do ETAP raters focus while rating oral performances?

Verbal reports are the verbalization of thoughts by research participants. Although verbal reports have been criticized for being subjective, their analyses provide explanatory results for research as long as they are analyzed in a systematic way and supported by other kinds of evidence. Therefore, this methodology has been used widely by other researchers to figure out rater behaviors and their focuses while rating performances (Weigle, 1999). Similarly, verbal reports shed light on ETAP raters' focuses in rating for the current thesis. In order to answer to the last research question, 55 verbal reports of ETAP raters were analyzed through NVivo 12. As a result, six common patterns were formed out of the transcribed data. These are (1) accent and sharing the same mother tongue with examinees, (2) testing process and the interlocutor in the exam, (3) the rating scale, (4) criteria, (5) appropriateness and meaningfulness of answers and (6) comparison with other test takers and across tasks.

In terms of accent and having the same L1 with test takers, raters mostly pay attention to the legibility of the test takers' accents. They do not think having a Turkish accent creates any conflict as long as it is clearly understandable by other speakers of English. Therefore, they signify whether the examinees' accents are internationally legible or not.

Aviation English possesses a limited domain and is only utilized for particular purposes for interaction in aviation. It is an unchanging restricted language and should be learnt by the native speakers of English as well. Therefore, it has been spoken as a lingua franca by the ones who have no shared language. At that point, mutual intelligibility plays a key role in communication (Estival, Farris, &

Molesworth, 2016). Likewise, Seiler (2009) also stresses the role of intelligibility from the point of raters by saying:

> A good rater is one who recognizes what features of someone's English are likely to give rise to problems of intelligibility. Ideally, a rater will therefore be familiar with all the major varieties of English and will have an understanding of phonetics and phenomena such as L1 interference. The question of whether the rater is an English native speaker is immaterial; what is crucial is that he or she can consistently assess speech samples and rate them appropriately. (p. 45)

That is to say, being a nonnative speaker of English and sharing the same mother tongue with test takers will not create any problem as long as raters focus on intelligibility between interlocutors during the rating process. With the same understanding that Seiler (2009) emphasizes, ETAP raters also pay attention to the international intelligibility of examinees' accents by taking their shared L1 into consideration.

Secondly, ETAP raters commented on the testing process and the interlocutor in their retrospective verbal reports. They think that the instructions of the questions should be provided to the examinees clearly and there should be a time limitation for test takers' answers so that raters can evaluate equal amount of linguistic output. Raters also query the repetition of questions by the interlocutor. They reckon that there should be a standardization regarding the repetition of the questions for examinees to have the same amount of chance to answer them. Moreover, raters emphasize that interlocutors in the tests are supposed to utilize the target language flawlessly since test takers can be affected by examiners' pronunciation mistakes during the exam and this may influence their scores.

Alderson (2011) recommends that it is crucial to have ongoing monitoring on the process of aviation English testing. Fortunately, raters are the ones who monitor all the rating process. Thus, they are significant participants of assessment process since their comments on that process are highly valuable to improve it. That is to say,

ETAP raters' comments on the testing process should be taken into consideration by the test holders so that the test can be improved to be more trustworthy.

Thirdly, raters refer to ETAP rating scale in their verbal reports. They claim that structure and vocabulary should be assessed separately in Task 4 and comprehension should be evaluated in the same task as well. Raters want to separate the criterion "Structure/Vocabulary" because they reckon that performance levels of examinees in both criteria can alter so a fair score might not be assigned for each criterion. Besides, examinees are expected to understand the given cases so that they can ask three related questions. When they do not understand the incidents, they cannot ask proper questions regarding the situation. This creates a problem for raters since they cannot take points off comprehension. Furthermore, Deniz and Olcay imply a need for a new middle category between descriptors in the scale. They do not want to assign neither four nor five to a performance but they think that there can be a middle point like four plus (+). This may indicate a problem with the categories in the rating scale. Koh (2003) and Orr (2002) (as cited in Ang-Aw & Chuen Meng Goh, 2011) assert that the ambiguity in the descriptors may cause difficulty in distinguishability between categories of a criterion. Thus, the feeling that a certain point is too low or too high for a performance can appear during the rating process. To overcome the problem, wording can be changed in those categories or a new middle category can be added into the scale. However, the functionality of the rating scale should be investigated with more data via Facet analyses. Finally, as for ETAP rating scale, all raters refer to it when assessing the performances especially when they are indecisive about the scores they assign. Brown (2003) claims that raters can evaluate different test takers with different levels of support due to their own way of rating which, in turn, risks the reliability of the assessment. In the assessment process, ETAP raters'

reference to the rating scale may hinder the huge different levels of evaluation and help to standardize the way of rating among raters.

As found in the previous studies (Ang-Aw & Chuen Meng Goh, 2011; Bogorevich, 2018), both criterion related and unrelated comments exist in ETAP raters' verbal reports. The relatedness of the comments regarding the criteria is decided by taking the descriptions in ETAP rating scale into consideration. Results show that 95% of raters' comments regarding criteria is related to the descriptors in the ETAP rating scale. That is to say, since the raters follow the descriptors consistently, their rating behaviors are consistent and they are reliable in rating. At this point, it seems that the results of qualitative analyses are in line with the findings of MFRM analysis indicating that raters are consistent in rating. On the other side, nine criterion irrelevant codes were generated out of the transcribed data. It is found that murmuring was an irrelevant remark for pronunciation. Hesitations or fluency, L1 influence and low frequency in usage were unrelated to the criterion structure in ETAP scale. There is no focus on pronunciation mistakes in vocabulary usage but raters mentioned it. Slip of tongue was an inappropriate consideration for fluency and clarification questions do not belong to the criterion comprehension. In terms of assessing interaction, comments regarding hesitations and understanding the questions were irrelevant. Upon this issue, Brown (2000) and Douglas (1994) (as cited in Ang-Aw & Chuen Meng Goh, 2011) state that the construct validity of the test can be risked if raters focus on criterion irrelevant factors which they think significant. Although ETAP raters have certain unrelated comments on criteria, it is found that they mostly follow the descriptors in ETAP rating scale. Therefore, the detailed findings can be utilized as individual feedback for each rater in rater training sessions to maintain the construct validity of the test. Furthermore, the impacts of

certain criteria on others were reported in the oral recordings. Raters deem that the low performances in structure and vocabulary affect fluency, comprehension and interaction adversely. They also claim that there are effects of structure on vocabulary usage and influence of fluency on structure. In order to figure out the impacts of some criteria on others, the performances of the test takers should be analyzed linguistically in the future studies.

Test takers are expected to ask appropriate questions for the given situations in Task 4 and they need to answer the follow up questions properly in the other tasks. Therefore, ETAP raters are supposed to assess the appropriateness and meaningfulness of the examinees' answers as well. Raters' reports show that they are not always certain about the appropriateness and meaningfulness of the answers because of their unfamiliarity with the aviation field. Related to this argument, Ang-Aw and Chuen Meng Goh (2011) state that the construct to be assessed in oral examinations can be operationalized by raters during the assessment process, which may have critical implications on the reliability of the test. In order to prevent this, more information regarding the proper answers for specific questions and a sample answer key can be provided to the raters before the rating sessions so that they can judge the appropriateness and meaningfulness of the answers in a fair way.

Finally, qualitative analyses reveal that comparison with other test takers and across tasks exist among ETAP raters although it is not common. According to the raters, they do the comparison among examinees so that they can assign fair scores. However, norm referenced tests include comparison of examinees' performances while criterion reference tests contain rating scale descriptors for assessment (Ang-Aw & Chuen Meng Goh, 2011). ETAP is a criterion reference test so comparison is not a convenient way to evaluate different examinees' performances in ETAP. Thus,

avoidance of inter-examinee comparison should be reminded in ETAP rater training sessions. Additionally, raters can follow their own criteria like comparing the examinees' performances to assign scores when they cannot find any reference for the performances in the rating scale. Therefore, it should be considered that such inter-examinee comparison can be resulted from the vagueness in the descriptors (Koh, 2003; Orr, 2002, as cited in Ang-Aw & Chuen Meng Goh, 2011). However, as stated before, more data is required to run a facet analysis for the rating scale functionality. Moreover, raters do the inter-task comparison within the same examinee's performance. That is to say, they compare the same criterion in different tasks for the same test taker. However, it should be reminded that there are minor differences in the descriptions of the same criterion in the rating scale of the different tasks since each task aims certain construct to be assessed.

5.6 Conclusion

This chapter presents a thorough discussion regarding the findings of the research by referring to the previous studies in the field. Upon this discussion, research implications, limitations and suggestions for future studies are addressed in the next chapter.

CHAPTER 6

CONCLUSION

6.1 Introduction

This study aims to investigate rater effects and focuses in ETAP conducted by PILVAK collaborating with BUYEM. ETAP raters were provided with online rater training for eight weeks. In the first three weeks, ICAO rating scales and principals were introduced to the raters and they evaluated previously graded twenty-two short speaking performances. For the last five weeks of the training, ETAP task specifications and rating scale were introduced. During those sessions, raters graded eight different test takers who had taken ETAP. Four raters out of eight participated in the current study. The scores that they assigned to the examinees and their verbal reports recoded after their evaluation constituted the data to be analyzed by means of Minifac (limited Rasch analysis software) and NVivo 12. Through Rasch analysis, ETAP raters' individual level of severity, consistency, central tendency and bias against criteria were examined while codes and patterns were created to find out raters' focuses and common comments during their rating process by utilizing NVivo 12 to analyze 150 minutes long verbal reports. The findings of the analyses are presented below.

Firstly, the severity levels of ETAP raters were investigated by checking Wright map, separation and reliability statistics in raters' measurement report revealed by Minifact. It is found that each rater belongs a different level of severity so they are not interchangeable. Weigle (1998) claims that raters may exhibit different severity levels even after rater training sessions and Eckes (2009) states that differences in severity levels are not a problem if raters are consistent in their rating

behavior. Therefore, it is assumed that different severity levels do not cause any problem as long as ETAP raters are consistent in their rating.

Secondly, ETAP raters' consistency in rating was examined by reviewing the fit statistics in Facets. The statistical results demonstrate that ETAP raters are consistent in the way they rate the examinees so they are reliable in rating.

Thirdly, whether a central tendency effect exist among ETAP raters were questioned to understand the even usage of categories in the ETAP rating scale. Raters' fit statistics, two different partial credit models for both raters and criteria and examinees' separation statistics were examined to find out the central tendency effect among raters. While raters' fit statistics and the examinees' separation statistics indicate absence of central tendency effect, partial credit models present more detailed analyses and illustrate some statistical evidence for existence of central tendency effect. In terms of criteria usage, the first partial credit model shows a central tendency effect for the criterion pronunciation in Task 5. Furthermore, the second partial credit model for each raters indicate that one rater may exhibit a central tendency effect for certain criteria (vocabulary, pronunciation, comprehension, interaction and content). This can be due to the monitoring effect on the rater (Myford & Mislevy, 1995, as cited in Knoch et al., 2007). Since s/he knew that s/he was observed during the assessment process, s/he might want to avoid making mistakes and utilized the middle categories at most. However, more performances should be assessed by the same rater to obtain a better interpretation regarding a central tendency. Besides, except for one rater, ETAP raters did not display any central tendency effect, which, in turn, may imply the elaborative descriptions of the categories in ETAP rating scale or the benefit of the online

training which results in less central tendency among raters (Wolfe & McVay, 2010). Yet, future research designs are required to make sure these implications.

Fourthly, ETAP raters' bias against any criteria was analyzed by conducting 'Rater x Criteria' bias/interaction analysis. The bias/Interaction table does not show any bias report whereas the pairwise comparison analyses illustrate certain bias patterns for raters. In terms of inter-rater comparison, there are significant differences in ratings between raters. Two raters rated the criterion "Structure/Vocabulary" in Task 4 more severely than the other two raters. As for the intra-rater comparison, two raters exhibit different individual severity levels when they rate the criteria "Structure/Vocabulary" in Task 4 and "Interaction/Content" in Task 7. The raters might consider certain criteria as quite significant for certain tasks so they might display bias against those criteria (Eckes, 2012). The results of the bias analyses can be utilized to detect the patterns in raters' rating behaviors and they can be used as feedback in rater training sessions. Furthermore, fair scores for test takers can be attained via facet analyses in spite of rater bias exhibited in the assessment.

Finally, ETAP raters' focuses when assessing oral performances were investigated by analyzing the verbal reports that were recorded by the raters themselves after assigning scores to the performances. Six prevalent points were detected based upon codes and patterns out of the data: (1) accent and sharing the same mother tongue with examinees, (2) testing process and the interlocutor in the exam, (3) the rating scale, (4) criteria, (5) appropriateness and meaningfulness of answers and (6) comparison with other test takers and across tasks. It is obvious that the qualitative data analyses shed light on the assessment process along with the quantitative one. Therefore, each of these focuses can be used as feedback for both

the raters and the test holders in order to develop the test, testing process and the rating scale to be more reliable and acceptable to a great extent.

To conclude, the results of the study reveal certain evidence for reliability of ETAP raters in oral performance assessment. The findings can be utilized for the improvement of raters and the assessment process as well. However, it should be reminded that this is a pioneer and tentative study concerning ETAP raters with small number of participants and test takers. Therefore, a replication of the study is required with larger number of raters and examinees in order to draw more valid conclusions.

6.2 Research implications

In ESP assessment, from developing the test to assigning scores, the whole process should be handled meticulously. When it comes to aviation English tests for aviation personnel, an elaborate assessment process should be ensured due to safety concerns in the sky. However, a common concern regarding the confidence in the quality of the assessment in aviation English tests continues. In terms of rating process, there are cases in which no respond is gained for rater reliability from aviation English test centers (Alderson, 2011). Based on this necessity for air safety, language test holders in aviation industry should declare their statistical reports so that test takers can ensure that fair scores are assigned by raters. Moreover, with the ongoing and detailed research concerning the assessment process, tests can gain worldwide trustworthiness.

To conduct a continuing and elaborative monitoring on raters, individual rater feedback in rater training sessions plays a significant role due to the particular differences among raters. Thus, in training sessions, results of facet analyses can be

utilized as individual feedback for raters. Rater trainers can develop a special program for them to understand rating scales in depth and to apply the scales both properly and consistently (Myford & Wolfe, 2004). Moreover, the qualitative analyses on the raters' verbal reports can enhance the quality of the rater training sessions by concentrating on the diverse needs of each rater.

Lastly, the application of MFRM model should become widespread in aviation English testing centers since it is beneficial to identify rater effects on scores and to assure validity of performance assessment. For this reason, it is recommended that specialist in the field of language assessment should be aware of the theories in psychometrics and know how to practice statistical analyses so that they can provide both feedback to raters and fair scores to test takers (Eckes, 2009). Besides, assessment validity can be improved with a mixed methods study since both qualitative and quantitative analyses complete each other despite being time consuming (Weigle, 1999).

6.3 Limitations and suggestions for future research

One major limitation of this research was the small number of raters who participated in the study and test takers who were rated for rater standardization in scoring. Therefore, this study provides preliminary and tentative results due to the restricted number of participants. The MFRM and qualitative analyses could have been conducted with a much larger number of raters and test takers in order to draw stronger conclusions on rater effects and raters' rating behaviors in an oral performance assessment of an aviation English test.

Another limitation of the study was its restricted research scope including only rater effects and focuses. With a larger number of participants, the functionality

of ETAP rating scale and item difficulty regarding each ETAP task can be examined through facet analysis for the future studies. Furthermore, test takers' previous scores in high stakes tests like TOEFL, IELTS or other aviation English tests can be compared with their scores in ETAP to investigate the concurrent validity of the test. As for the qualitative part of the study, triangulation of data could have been generated with the collection of more data via interviews and online surveys for background information.

For the future studies, ETAP raters' individual differences in rating and the factors for the differences can be investigated with more detailed verbal reports and interviews. Through both qualitative and descriptive analyses (MFRM results), raters can be provided with individual feedback and the impacts of feedback on rating behaviors can be examined throughout rater training process.

In conclusion, this research shows that a multidisciplinary study can be conducted by paying attention to the primary considerations in language assessment, aviation industry, Rasch measurement and rater effects. In addition to the consequential results of this study, which presents benefits for oral performance assessment in ESP, it also shows that assessment process of aviation English tests should be monitored constantly and elaborately for air safety.

APPENDIX A

ETAP RATING SCALES

Table A1. ETAP Task 4

LEVEL	STRUCTURE/VOCABULARY
Expert 6	Question structures and vocabulary are precise and elaborate. Questions are formulated in order to get detailed and purposeful responses.
Extended 5	The speaker can use complex question structures with only few negligible errors. Vocabulary choice is precise. Questions are focused on the topic.
Operational 4	Syntactic structures of the questions are correct except for minor mistakes. Vocabulary is appropriate although it can be slightly imprecise. Almost all questions are relevant. Listener understands the questions.
Pre-Operational 3	Syntactic errors in question structures. Generally incorrect choice of vocabulary, or the candidate experiences difficulty in finding vocabulary. Questions are generally not relevant or too general. Listener may not be able to understand the meaning at times.
Elementary 2	Performs at a level below the pre-operational level.
Pre-elementary 1	Performs at a level below the Elementary level. There is not enough output to evaluate.

Table A2. ETAP Task 5

LEVEL	PRONUNCIATION Assumes a dialect and/or accent intelligible to the aeronautical community.	STRUCTURE Relevant grammatical structures and sentence patterns are determined by language functions appropriate to the task.	VOCABULARY	FLUENCY	INTERACTION/CONTENT
Expert 6	Pronunciation, stress, rhythm and intonation, though possibly influenced by the first language or regional variation, almost never interfere with ease of understanding.	Both basic and complex grammatical structures and sentence patterns are consistently well controlled.	Vocabulary range and accuracy are sufficient to communicate effectively on a wide variety of familiar and unfamiliar topics. Vocabulary is idiomatic, nuanced, and sensitive to register.	Able to speak at length with a natural, effortless flow. Varies speech flow for stylistic effect, e.g. to emphasize a point. Uses appropriate discourse markers and connectors spontaneously.	Can form a coherent and relevant narration of events with elaborate details and logical interpretation. The information provided in follow-up questions are extensive, shows the ability of the speaker to speculate on issues. Interacts with ease.
Extended 5	Pronunciation, stress, rhythm, and intonation, though influenced by the first language or regional variation, rarely interfere with ease of understanding.	Basic grammatical structures and sentence patterns are consistently well controlled. Complex structures are attempted but with errors which sometimes interfere with meaning.	Vocabulary range and accuracy are sufficient to communicate effectively on common, concrete, and work-related topics. Paraphrases consistently and successfully. Vocabulary is sometimes idiomatic.	Able to speak at length with relative ease on familiar topics but may not vary speech flow as a stylistic device. Can make use of appropriate discourse markers or connectors.	Coherent and relevant description of events with slight lapses and diversions. Adequate details are provided to build up a whole narrative. The speaker attempts at speculation and provides relevant details in the follow-up questions. Interacts successfully.

Operational 4	Pronunciation, stress, rhythm, and intonation are influenced by the first language or regional variation but only sometimes interfere with ease of understanding.	Basic grammatical structures and sentence patterns are used creatively and are usually well controlled. Errors may occur, particularly in unusual or unexpected circumstances, but rarely interfere with meaning.	Vocabulary range and accuracy are usually sufficient to communicate effectively on common, concrete, and work-related topics. Can often paraphrase successfully when lacking vocabulary in unusual or unexpected circumstances.	Produces stretches of language at an appropriate tempo. There may be occasional loss of fluency on transition from rehearsed or formulaic speech to spontaneous interaction, but this does not prevent effective communication. Can make limited use of discourse markers or connectors. Fillers are not distracting.	Basic events are correctly interpreted and put in a narrative frame that can be followed without loss of meaning. May lack elaboration or detailed interpretation but provides relevant information in the follow-up questions. Can manage interaction in the cases of misunderstandings by checking, confirming, or clarifying.
Pre-Operational 3	Pronunciation, stress, rhythm and intonation, are influenced by the first language or regional variation, and frequently interfere with ease of understanding.	Basic and grammatical structures and sentence patterns associated with predictable situations are not always well controlled. Errors frequently interfere with meaning.	Vocabulary range and accuracy are often sufficient to communicate on common, concrete, or work-related topics, but range is limited and the word choice often inappropriate. Is often unable to paraphrase successfully when lacking vocabulary.	Produces stretches of language, but phrasing and pausing are often inappropriate. Hesitations or slowness in language processing may prevent effective communication. Fillers are sometimes distracting.	Inconsistent and sporadic narration of events. Necessary details are not provided while minor details might be mentioned. The listener can get confused. The answers to follow-up questions are uninformative and fragmentary. May not be able to deal with communication problems successfully.

Elementary 2	Pronunciation, stress, rhythm, and intonation are heavily influenced by the first language or regional variation and usually interfere with ease of understanding.	Shows only limited control of a few simple memorized grammatical structures and sentence patterns.	Limited vocabulary Range and consisting only of isolated words and memorized phrases.	Can produce very short, isolated, memorized utterances with frequent pausing and a distracting use of fillers to search for expressions and to articulate less familiar words.	Performs at a level below the pre-operational level.
Pre-elementary 1	Performs at a level below the Elementary level.	Performs at a level below the Elementary level.	Performs at a level below the Elementary level.	Performs at a level below the Elementary level.	Performs at a level below the Elementary level.

Table A3. ETAP Task 6 and 7

LEVEL	PRONUNCIATION Assumes a dialect and/or accent intelligible to the aeronautical community.	STRUCTURE Relevant grammatical structures and sentence patterns are determined by language functions appropriate to the task.	VOCABULARY	FLUECNY	COMPREHENSION	INTERACTION/CONTENT
Expert 6	Pronunciation, stress, rhythm and intonation, though possibly influenced by the first language or regional variation, almost never interfere with ease of understanding.	Both basic and complex grammatical structures and sentence patterns are consistently well controlled.	Vocabulary range and accuracy are sufficient to communicate effectively on a wide variety of familiar and unfamiliar topics. Vocabulary is idiomatic, nuanced, and sensitive to register.	Able to speak at length with a natural, effortless flow. Varies speech flow for stylistic effect, e.g. to emphasize a point. Uses appropriate discourse markers and connectors spontaneously.	Full comprehension of main points and important details; clear understanding of importance of critical events and relations between events/phenomena/objects; inferring this when not explicitly stated.	Can form a thorough and coherent summary of the text including major details by showing relations such as comparison and contrast, reason and result, etc. Paraphrases successfully when needed. The information provided in follow-up questions are extensive, shows the ability of the speaker to speculate on issues. Interacts with ease.

Extended 5	Pronunciation, stress, rhythm, and intonation, though influenced by the first language or regional variation, rarely interfere with ease of understanding.	Basic grammatical structures and sentence patterns are consistently well controlled. Complex structures are attempted but with errors which sometimes interfere with meaning.	Vocabulary range and accuracy are sufficient to communicate effectively on common, concrete, and work-related topics. Paraphrases consistently and successfully. Vocabulary is sometimes idiomatic.	Able to speak at length with relative ease on familiar topics but may not vary speech flow as a stylistic device. Can make use of appropriate discourse markers or connectors.	Comprehension of main ideas with some details; most critical events are identified. Unstated critical relations between events/phenomena/objects may not be inferred fully.	Can form a coherent summary of the text including most of the main ideas. There might be unimportant lapses and omissions but the speaker can establish the logical relations between the important points in the text. The speaker attempts at speculation and provides relevant details in the follow-up questions. Interacts successfully.
Operational 4	Pronunciation, stress, rhythm, and intonation are influenced by the first language or regional variation but only sometimes interfere with ease of understanding.	Basic grammatical structures and sentence patterns are used creatively and are usually well controlled. Errors may occur, particularly in unusual or unexpected circumstances, but rarely interfere with meaning.	Vocabulary range and accuracy are usually sufficient to communicate effectively on common, concrete, and work-related topics. Can often paraphrase successfully when lacking vocabulary in unusual or unexpected circumstances.	Produces stretches of language at an appropriate tempo. There may be occasional loss of fluency on transition from rehearsed or formulaic speech to spontaneous interaction, but this does not prevent effective communication. Can make limited use of discourse markers or connectors. Fillers are not distracting	Comprehension of most main points but may skip some important details. Comprehension may be limited to clearly stated facts in the text. Situationally complex cases may not be understood clearly with all aspects when the text is complex.	The summary of the text includes most of the basic points in the text and reported in a coherent manner and can be followed without loss of meaning. Can paraphrase but into simpler forms. May lack elaboration or detailed interpretation but provides relevant information in the follow-up questions. Can manage interaction in the cases of misunderstandings by checking, confirming, or clarifying.

Pre-Operational 3	Pronunciation, stress, rhythm and intonation, are influenced by the first language or regional variation, and frequently interfere with ease of understanding.	Basic and grammatical structures and sentence patterns associated with predictable situations are not always well controlled. Errors frequently interfere with meaning.	Vocabulary range and accuracy are often sufficient to communicate on common, concrete, or work-related topics, but range is limited and the word choice often inappropriate. Is often unable to paraphrase successfully when lacking vocabulary.	Produces stretches of language, but phrasing and pausing are often inappropriate. Hesitations or slowness in language processing may prevent effective communication. Fillers are sometimes distracting.	Can understand some of the basic points or clearly stated facts but cannot follow or connect most of the ideas in the text. Comprehension is limited to familiar topics.	Inconsistent and sporadic reporting of the events/phenomena/objects in the text. Reports what is heard rather than important points. Unsuccessful or incorrect paraphrase. The listener can get confused. The answers to follow-up questions are uninformative and fragmentary. May not be able to deal with communication problems successfully.
Elementary 2	Pronunciation, stress, rhythm, and intonation are heavily influenced by the first language or regional variation and usually interfere with ease of understanding.	Shows only limited control of a few simple memorized grammatical structures and sentence patterns.	Limited vocabulary Range and consisting only of isolated words and memorized phrases.	Can produce very short, isolated, memorized utterances with frequent pausing and a distracting use of fillers to search for expressions and to articulate less familiar words.	Performs at a level below the pre-operational level.	Performs at a level below the pre-operational level.
Pre-elementary 1	Performs at a level below the Elementary level.	Performs at a level below the Elementary level.	Performs at a level below the Elementary level.	Performs at a level below the Elementary level.	Performs at a level below the Elementary level.	Performs at a level below the Elementary level.

REFERENCES

Aiguo, W. (2007). Teaching aviation English in the Chinese context: Developing ESP theory in a non-English speaking country. *Science Direct,* 26, 121-128.

Aiguo, W. (2008). Reassessing the position of aviation English: from a special language to English for specific purposes. *Ibérica, 15*, 151-163.

Alderson, J. C. (2009). Air safety, language assessment policy, and policy implementation: The case of aviation English. *Annual Review of Applied Linguistics*, *29*, 168–187. doi: 10.1017/s0267190509090138

Alderson, J. C. (2010). A survey of aviation English tests. *Language Testing, 27*(1), 51-72.

Alderson, J. C. (2011). The Politics of Aviation English Testing. *Language Assessment Quarterly, 8*(4), 386-403. doi:10.1080/15434303.2011.622017

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. (1991). What does language testing have to offer? *TESOL Quarterly*, *25*, 671–704.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19*(4), 453-476.

Barkaoui, K. (2013). Multifaceted Rasch analysis for test evaluation. *The Companion to Language Assessment,* 1-46. doi:10.1002/9781118411360.wbcla070

Bogorevich, V. (2018). *Native and non-native raters of L2 speaking performance: Accent familiarity and cognitive process.* (Unpublished doctoral thesis). Northern Arizona University, San Francisco.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Dordrecht: Springer Netherlands.

Brown, A. D. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12*(1), 1–16.

Brown, A. (2003) Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20*(1), 1-25.

Bullock, N. (2017). A re-evaluation of washback for learning and testing language in aeronautical communications. *International Civil Aviation English Association Workshop (ICAEA), 19*, 1-30.

Cheng, W. (2010). Hong Kong engineering corpus: Empowering professionals-in-training to learn the language of their profession. In M. C. Campoy-Cubillo, B. Bell´es-Fortu˜no, & M. L. Gea-Valor (Eds.), *Corpus-based approaches to English language teaching* (pp. 67–78). London/New York: Continuum.

Cushing, S. (1994). *Fatal words: Communication clashes and aircraft crashes*. Chicago, IL: The University of Chicago Press.

Davies, A. (2001). The logic of testing languages for specific purposes. *Language Testing, 18*, 133–147.

Day, B. (2004). Heightened awareness of communication pitfalls can benefit safety. *ICAO Journal, 59*, 20-22.

Douglas, D. (2000). *Assessing languages for specific purposes*: Cambridge University Press.

Douglas, D. (2001). Language for Specific Purposes assessment criteria: where do they come from? *Language Testing*, *18*(2), 171-185.

Douglas, D. (2004). English language testing in the context of Aviation English. *ICAO Journal, 59*(3), 17-18.

Du, Y., Wright, B. D., & Brown, W. L. (1996, April). *Differential facet functioning detection in direct writing assessment*. Paper presented at the Annual Conference of the American Educational Research Association, New York, NY.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A Many-Facet Rasch analysis. *Language Assessment Quarterly, 2*(3), 197-221. doi:10.1207/s15434311laq0203_2

Eckes, T. (2009). Many-facet Rasch measurement. *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*. (Section H). Strasbourg, France: Council of Europe/Language Policy Division. Retrieved from http://www.coe.int/t/dg4/Linguistic/CEF-refSupp-SectionH.pdf

Elder, C. (2001). Assessing the language proficiency of teachers: Are there any border controls? *Language Testing, 18*(2), 149–170.

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly, 2*(3), 175–196.

Emery, H., J. (2014) Developments in LSP Testing 30 Years On? The Case of Aviation English. *Language Assessment Quarterly, 11*(2), 198-215, DOI: 10.1080/15434303.2014.894516

Estival, D., Farris, C., & Molesworth, B. (2016). *Aviation English: a Lingua Franca for pilots and air traffic controllers*. Routledge, Taylor & Francis Group.

Flowerdew, J. (1997). Corpus linguistics: Applications to ESP. Hong Kong University of Science & Technology, Working papers from the Language Centre: May 1997, 100–108. Retrieved April 10, 2012, from http://repository.ust.hk/dspace/bitstream/1783 1/1377/1/explorelang08.pdf

Hawkey, R. (1978). *English for special purposes*. London: British Council English Teaching Centre.

Heigham, J., & Croker, R. A. (2009). *Qualitative research in applied linguistics*. Palgrave Macmillan.

Hinrich, S. W. (2008). *The Use of Questions in International Pilot and Air-traffic Controller Communication*: ProQuest. Retrieved from https://shareok.org/bitstream/handle/11244/7069/ English%20Department_20.pdf?sequence=1.

Howard, J. W. (2008). "Tower, am I cleared to land?": Problematic communication in aviation discourse. *Human Communication Research, 34*(3), 370-391.

Hutchinson, T., & Waters, A. (1987). *English for Specific Purposes: A learning-centred approach.* Cambridge: Cambridge University Press.

International Civil Aviation Organization. (2001). Annex 10. *International Standards and Recommended Practices, Aeronautical Telecommunications: International Civil Aviation Organization, 1*.

International Civil Aviation Organization. (2004). *Manual on the implementation of ICAO language proficiency requirements*. ICAO Doc 9835 AN/453. International Civil Aviation Organization: Montreal, Quebec, Canada. Retrieved from http://caa.gateway.bg/upload/docs/9835_1_ed.pdf)

International Civil Aviation Organization. (2010). *Manual on the implementation of ICAO Language Proficiency Requirements* (2nd ed., Doc 9835 AN/453). Montreal, Canada: Author.

Jones, R. K. (2003). Miscommunication between pilots and air traffic control. *Language problems & language planning, 27*(3), 233-248.

Kachru, B. (2003). Liberation linguistics and the quirk concern. In Seidlhofer, B. (Ed.) *Controversies in Applied Linguistics* (pp. 19-33). Oxford, England: Oxford University Press.

Kasper, G. (1998). Analysing verbal protocols. *TESOL Quarterly, 32*(2), 358-362.

Kaygan, A. D. (2005). *Özel amaçlı İngilizce öğretiminde gereksinim çözümlemesi (hava trafik terminolojisi eğitimi alan pilotlar ile alan çalışması)*. (Unpublished master's thesis). Marmara University, Istanbul, Turkey.

Kim, H. (2012). *Exploring the construct of aviation communication: A critique of the ICAO language proficiency policy.* (Unpublished doctoral thesis). University of Melbourne, Australia.

Kim, H. J. (2015). A Qualitative Analysis of Rater Behavior on an L2 Speaking Assessment. *Language Assessment Quarterly, 12*(3), 239–261. doi: 10.1080/15434303.2015.1049353

Kim, H., & Elder, C. (2009). Understanding aviation English as a lingua franca: Perceptions of Korean aviation personnel. *Australian Review of Applied Linguistics, 32*(3), 23-23.

Kim, H., & Elder, C. (2015). Interrogating the construct of aviation English: Feedback from test takers in Korea. *Language Testing, 32*(2), 129-149.

Knoch, U. (2007). Do empirically developed rating scales function differently to conventional rating scales for academic writing? *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, 5, 1–36.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing, 12*, 26-43.

Knoch, U. (2009). Collaborating with ESP stakeholders in rating scale validation: The case of the ICAO rating scale. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 7*, 21–46.

Kuckartz, U. (2014). *Qualitative text analysis*. Sage Publications.

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

Küçük, F. (2017). *Assessing academic rating skills in Turkish as a foreign language*. (Unpublished master's thesis). Boğaziçi University, İstanbul.

Li, H., & He, L. (2015). A comparison of EFL raters' essay-rating processes across two types of rating scales. *Language Assessment Quarterly, 12*(2), 178–212. https://doi.org/10.1080/15434303.2015.1011738

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.

Linacre, J. M. (1989, March). *Rasch models from objectivity: A generalization*. Paper presented at the International Objective Measurement Workshop, Berkeley, CA.

Linacre, J. M. (1996). Generalizability theory and many-facet Rasch measurement. In Engelhard, G. & Wilson, M. (Eds.), *Objective measurement: Theory into practice, Volume 3* (pp. 85-98). Norwood, NJ: Ablex.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness, *Journal of Applied Measurement, 3*(1), 85-106.

Linacre, J. M. (2007). *A user's guide to FACETS Rasch model computer program.* Available online at: www.winstep.com

Linacre, J. M., & Wright, B. D. (2004). Constructing measures from many-facet data. In E. V. Smith, Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 296-321). Maple Grove, Minnesota: JAM Press.

Lumley, T. (1998). Perceptions of language-trained raters and occupational experts in a test of occupational English language proficiency. *English for Specific Purposes, 17,* 347–367.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54-71.

Lunz, M. E., Stahl, J. A.. Wright, B. D., & Linacre, J. M. (1989). *Variation among examiners and protocols on oral examinations.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

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

McNamara, T. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing, 7*(1), 52-76. doi:10.1177/026553229000700105

McNamara, T. F. (1996). *Measuring second language performance*. London, UK: Longman.

McNamara, T. (2012). *At last: Assessment and English as a lingua franca.* Plenary talk at 5th International Conference of English as a Lingua Franca, 24 May, Istanbul.

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing, 29*(4), 555-576. doi:10.1177/0265532211430367

Merriam, S. B., & Tisdell, E. J. (2015). *Qualitative research: a guide to design and implementation*. (4th ed.). Jossey-Bass.

Miles, M. B., Huberman, A. M., & Saldana, J. (2014). *Qualitative data analysis*. (3rd ed.). Sage Publications.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386-422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*(2), 189-227.

Norris, J., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). Designing second language performance assessments (Technical Report 18). Honolulu: University of Hawaii.

O'sullivan, B. (2012). Assessment Issues in Languages for Specific Purposes. *The Modern Language Journal, 96,* 71-88. doi:10.1111/j.1540-4781.2012.01298.x

O'Sullivan, B., & Weir, C. J. (2011) Language testing = validation. In B. O'Sullivan (Ed.), *Language testing theories and practices* (pp. 13–32). Oxford: Palgrave Macmillan.

Paltridge, B., & Starfield, S. (2013). *The handbook of English for specific purposes*. Chichester etc.: Wiley-Blackwell.

Park, M. (2015). *Development and validation of virtual interactive tasks for an aviation English assessment*. (Unpublished doctoral thesis). Iowa State University.

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, Cambridge and Arnhem* (pp. 74–91). Cambridge, UK: Cambridge University Press.

Prinzo, O. V., & Thomson, A. C. (2009). *The ICAO English Language Proficiency Rating Scale applied to enroute voice communications of U.S. and foreign pilots* (DOT/FAA/AM-09/10). Washington, DC: Federal Aviation Administration. Retrieved from www.dtic.mil/dtic/tr/fulltext/u2/a500318.pdf

Ripley, R. F., & Finch, J. L. (2004). The efficacy of standard aviation English. In M. A. Turney (Ed.), *Tapping diverse talent in aviation: Culture, gender, and diversity* (pp. 99–103). Hampshire, England: Ashgate.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25*(4), 463-492.

Seiler, W. (2009). English as a lingua franca in aviation. *English Today*, *25*(2), 43–48. https://doi.org/10.1017/s0266078409000182

Shirazi, M. A. (2019). For a Greater Good: Bias Analysis in Writing Assessment. *SAGE Open*, *9*(1), 1-14. doi:10.1177/2158244018822377

Sudweeks, R., Reeve, S., & Bradshaw, W. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*, 239-261.

Swales, J. (1985). *Episodes in ESP: a source and reference book on the development of English for science and technology.* Oxford: Pergamon Press.

Trippe, J. E. (2018). *Aviation English is distinct from conversational English: Evidence from prosodic analyses and listening performance*. (Unpublished doctoral thesis). University of Oregon.

Ünaldı, A. (2019). *Validation document for the English test for aviation personnel.* Unpublished manuscript.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263–287.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, *6*(2), 145–178.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* Houndmills, England: Palgrave Macmillan.

Wilson, M., & Draney, K. (2000, May). *Standard mapping: A technique for setting standards and maintaining them over time*. Paper presented at the international conference on measurement and multivariate analysis, Banff, Canada.

Wolfe, E. W., & McVay, A (2010). *Rater effects as a function of rater training context*. New York: Pearson Research and Innovation Network.

Woodward-Kron, R., & Elder, C. (2016). A comparative discourse study of simulated clinical roleplays in two assessment contexts: Validating a specific purpose-language test. *Language Testing, 33*(2), 251-270. doi:10.1177/0265532215607399

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.

Yılmaz, F. (2017). Analysis of the rater effects in the rating of diagnostic trees prepared by teacher candidates by the many-facet rasch model. *Journal of Education and Practice, 8*(18), 174-184.