A STUDY OF RATERS' BACKGROUND KNOWLEDGE, RATER TRAINING AND OTHER FACTORS AFFECTING THEIR DECISION MAKING IN RATING THAI PILOTS' ENGLISH SPEAKING PROFICIENCY

Acting Sub-lieutenant Sutas Dejkunjorn

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy Program in English as an International Language
Graduate School
Chulalongkorn University
Academic Year 2010

การศึกษาภูมิหลัง การฝึกอบรมการประเมิน และองค์ประกอบอื่นๆที่มีผลในการตัดสินใจของผู้ประเมิน ในการประเมินผลความสามารถในการพูดภาษาอังกฤษของนักบินไทย

ว่าที่ร้อยตรี สุทรรศน์ เดชกุญชร

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาศิลปศาสตรดุษฎีบัณฑิต
สาขาวิชาภาษาอังกฤษเป็นภาษานานาชาติ
บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2553

Thesis Title	A STUDY OF RATERS' BACKGROUND KNOWLEDGE, RATER TRAINING AND OTHER FACTORS AFFECTING THEIR DECISION MAKING IN RATING THAI PILOTS' ENGLISH SPEAKING PROFICIENCY
By	Acting Sub-lieutenant Sutas Dejkunjorn
Field of Study	English as an International Language
Thesis Advisor	Professor Kanchana Prapphal, Ph.D.

Accepted by the Graduate School, Chulalongkorn University in Partial Fulfillment of the Requirements for the Doctoral Degree

…….……Dean of the Graduate School

(Associate Professor Pornpote Piumsomboon, Ph.D.)

THESIS COMMITTEE

……..…. Chairperson

(Associate Professor Sumalee Chinokul, Ph.D.)

……..….Thesis Advisor

(Professor Kanchana Prapphal, Ph.D.)

……Examiner

(Associate Professor Prakaikaew Opanon-amata)

……Examiner

(Ajarn Tanyaporn Arya, Ph.D.)

……Examiner

(Ajarn Wannana Soontornnaruerangsee, Ph.D.)

สุทรรศน์ เดชกุญชร: การศึกษาภูมิหลัง การฝึกอบรมการประเมินและองค์ประกอบอื่นๆที่มีผลต่อการตัดสินใจของผู้ประเมินในการประเมินผลความสามารถในการพูดภาษาอังกฤษของนักบินไทย (A STUDY OF RATERS' BACKGROUND KNOWLEDGE, RATER TRAINING AND OTHER FACTORS AFFECTING THEIR DECISION MAKING IN RATING THAI PLOTS' ENGLISH SPEAKING PROFICIENCY) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ศ. ดร. กาญจนา ปราบพาล, 476 หน้า

การวิจัยนี้มีจุดประสงค์หลักเพื่อศึกษาผลกระทบของภูมิหลังของผู้ประเมิน การฝึกอบรมการประเมินของผู้ประเมิน และองค์ประกอบอื่นๆที่มีผลในการตัดสินใจของผู้ประเมินในการประเมินผลความสามารถในการพูดภาษาอังกฤษของนักบินไทย กลุ่มตัวอย่างในการศึกษาครั้งนี้ประกอบด้วยผู้ประเมินที่มีภูมิหลังทางภาษา และผู้ประเมินที่มีภูมิหลังทางการปฏิบัติการบิน จำนวนกลุ่มละสิบคนโดยการคัดเลือกแบบมีวัตถุประสงค์ ผู้วิจัยแบ่งกลุ่มผู้ประเมินออกเป็นสี่กลุ่มย่อยคือ กลุ่มที่มีภูมิหลังทางภาษาที่ผ่านการฝึกอบรมการประเมิน กลุ่มที่มีภูมิหลังทางภาษาที่ไม่ผ่านการฝึกอบรมการประเมิน กลุ่มที่มีภูมิหลังทางการปฏิบัติการบินที่ผ่านการฝึกอบรมการประเมิน และกลุ่มที่มีภูมิหลังทางการปฏิบัติการบินที่ไม่ผ่านการฝึกอบรมการประเมิน เครื่องมือที่ใช้ในการวิจัยได้แก่ แบบทดสอบความสามารถในการพูดภาษาอังกฤษสำหรับนักบิน (Pilot English speaking proficiency) ของ RELTA แบบสอบถามข้อมูลและความเห็นของผู้ประเมิน และการสัมภาษณ์แบบกึ่งควบคุม (Semi-structured interview) รูปแบบการวิจัยคือการวิจัยแบบกึ่งทดลอง (Quasi-experimental research) สถิติที่ใช้วิเคราะห์คือ การวิเคราะห์ความแปรปรวนแบบจำแนกสองทาง (2-way ANOVA) และการใช้สถิติที (*t-test*) เพื่อวิเคราะห์ข้อมูลเชิงปริมาณ นอกจากนี้ผู้วิจัยใช้การวิเคราะห์เชิงเนื้อหาวิเคราะห์ข้อมูลเชิงคุณภาพเพื่อให้ได้ข้อมูลเชิงลึกและเพื่อยืนยันผลของการวิเคราะห์เชิงปริมาณ ผลการวิจัยพบว่าภูมิหลังและการฝึกอบรมการประเมินที่แตกต่างกันไม่มีผลกระทบอย่างมีนัยสำคัญทางสถิติต่อการตัดสินใจในการประเมินผลความสามารถในการพูดภาษาอังกฤษของนักบินไทย อย่างไรก็ตาม ปัจจัยการฝึกอบรมการประเมินมีผลกระทบต่อตัวแปรตาม (Dependent variable) มากกว่าปัจจัยด้านภูมิหลังของผู้ประเมิน จากการวิจัยองค์ประกอบ 13 ประการที่อาจมีผลต่อการตัดสินใจของผู้ประเมิน ผลการวิจัยพบว่าสามารถแบ่งองค์ประกอบได้เป็นสามกลุ่ม คือ กลุ่มที่มีผลต่อการตัดสิน ได้แก่ กลยุทธ์ที่ใช้ในการประเมิน (Rating strategies) ตัวผู้เข้าสอบ (Test-takers) สเกลการให้คะแนนและรายละเอียดของสเกล (Rating scales & descriptors) ความสัมพันธ์ส่วนบุคคลระหว่างผู้ประเมินและผู้เข้าสอบ(Relationships) คะแนนตัดสินว่าผ่านหรือไม่ผ่าน (Cut-off score) วิธีการให้คะแนน(Scoring) กลุ่มที่ไม่มีผลต่อการตัดสินใจ ได้แก่ สภาพแวดล้อมทางกายภาพในการประเมิน (Physical settings) และ อิทธิพลของผู้สัมภาษณ์/คู่สนทนา (Interviewer effects) กลุ่มที่ไม่ชัดเจน ทำให้ไม่สามารถสรุปได้ ได้แก่ ภูมิหลังทางการศึกษาและภูมิหลังทางการประเมิน (Educational & Rating background) สภาพจิตใจ (Mental conditions) สภาพทางกายภาพ (Physical conditions) ลักษณะของแบบทดสอบและตัวอย่างการพูดของผู้เข้าสอบ (Test tasks & Speech samples) กลวิธีในการประเมิน (Rating strategies) และ ความเข้มงวดหรือการปล่อยคะแนนของผู้ประเมิน (Harshness/leniency)

สาขาวิชา ภาษาอังกฤษเป็นภาษานานาชาติ ลายมือชื่อนิสิต .....

ปีการศึกษา 2553 ลายมือชื่อ อ.ที่ปรึกษา วิทยานิพนธ์.....

4989706120: MAJOR ENGLISH AS AN INTERNATIONAL LANGUAGE
KEYWORDS: ENGLISH SPEAKING PROFICIENCY / RATERS' BACKGROUND / RATER TRAINING / FACTORS AFFECTING RATING / RATERS' DECISION-MAKING/ PILOT SPEAKING PROFICIENCY/ THAI PILOTS

SUTAS DEJKUNJORN: A STUDY OF RATERS' BACKGROUND, RATER TRAINING AND OTHER FACTORS AFFECTING THEIR DECISION MAKING IN RATING THAI PILOTS' ENGLISH SPEAKING PROFICIENCY. THESIS ADVISOR: PROF. KANCHANA PRAPPHAL, 476 pp.

The objectives of this study were to examine two kinds of raters having different background knowledge i.e. linguistic and operational raters with and without rater training experience when they assessed Thai pilots' English language speaking performances on RELTA, and to explore the other factors affecting their decision-making in awarding the scores to the candidates. The participants in the study were 20 raters. They were categorized into two main groups, linguistic and operational, based on their educational and professional background. Then, they were divided further into four sub-groups i.e. linguistic/trained, linguistic/untrained, operational/trained, and operational/untrained. The subjects participating in the main study were purposively selected for data analysis. The source of data was the RMIT English Language Test for Aviation (RELTA). The instruments were questionnaires, rater score sheet and remarks, and semi-structured interviews. The research design was the quasi-experimental research. The 2x2 ANOVA and t-test were used to analyze the quantitative data, and the content analysis was used to analyze the qualitative one, and to confirm the results obtained from the quantitative analysis.

It was found that both raters' background and rater training did not significantly affect the raters' decision-making in rating Thai pilots' English speaking proficiency, in both main and interaction effects. However, the factor of training seemed to affect more than the factor of background on the raters' rating scores. The content analysis analyzed 13 factors that might affect the raters' decision-making. The results revealed that they could be divided into three groups: the group of the factors which had effects on the raters' decision-making i.e. rating strategies, candidates/test-takers, rating scale and descriptors, personal relationships between raters and candidates, cut-off score, and scoring: the group of the factors which had no effect on the raters' decision-making i.e. physical settings, and interviewer/interlocutor: and the group of the factors which were not obvious, hence, unable to make a conclusion i.e. rater educational and rating background, rater mental conditions, rater's physical conditions, test tasks and speech samples, and raters' harshness/leniency.

Field of Study: English as an International Language Student's Signature .....

Academic Year : 2010 Advisor's Signature .....

ACKNOWLEDGEMENTS

This research study was motivated by the need to develop a proficiency test for Thai Airways International pilots in order to comply with the International Civil Aviation Organization requirements. It was kindly funded by the 90th Chulalongkorn University Anniversary Research Fund (Ratchadaphiseksomphot Endowment Fund). The researcher would not be able to finish this study without helps and supports from many people. Therefore, I would like to express my gratitude to the following research committee members for their important contributions. First, to my beloved advisor – Professor Kanchana Prapphal, Ph.D. for her endless support and guidance extended at various stages of this research and throughout my study. Then, to the Chair of the committee – Associate Professor Sumalee Chinokul, Ph.D. who also helped me get through some hard time and hard feeling during my study. In addition, to the other members of the committee i.e. Associate Professor Prakaikaew Opanon-amata; Ajarn Tanyaporn Arya, Ph.D.; and Ajarn Wannana Soontornnaruerangsee, Ph.D. for their valuable comments that helped strengthen the final draft of this research. I am also deeply indebted to Sutthinee Chuanchaisit, Ph.D. who is my classmate for her support, suggestions and encouragement in every way throughout my study. Without her, I would not be able to finish both this research and my study.

My special recognition also goes to Michael Kay who is the International Development Manager – Aviation of RMIT English Worldwide, and a very good friend of mine, who helped get the permission to use RELTA as a crucial instrument in this research. Furthermore, I would like to convey my appreciation to all 20 raters who gave their precious time to participate in the rating and the interview process. Some of them are my classmates. Some are my colleagues. Some are my friends. No matter who they are, all of them played an important role in making this research a successful one.

I particularly wish to express my gratefulness to my parents i.e. my father, who happens to be my lifetime English teacher, Police General (Retired) Vasit Dejkunjorn, Ph.D. (Honorary) and my mother, Khunying Tasna Bunnag Dejkunjorn. Without them, I would not be able to study anything in this world. Lastly, to my wife, Vivian Dejkunjorn, for her support, advice, and initiation in the beginning to study at this great institution - Chulalongkorn University.

CONTENTS

List of Tables

List of Figures

Figure Page

CHAPTER I
INTRODUCTION

1.1 Background of the study

English is regarded as a global language (Crystal, 1997). It acts as a "lingua franca" or "common language" when people speaking different languages try to communicate with each other. The language used in the world of aviation is no exception.

The prime objective of using a language in aviation context is safety. English is internationally exploited as a means to communicate among pilots and air traffic controllers to accomplish that crucial objective. International Civil Aviation Organization (ICAO), which is an agency under the United Nations (UN) and is responsible for the aviation safety, states in its Annex 10 that English is recommended to be made available whenever an aircraft station is unable to communicate in the language used by the station on ground (ICA0, 2004).

The most important problem is that not everyone involved in this context is proficient in English. The lack of this proficiency may lead to miscommunication in air traffic, which may induce an accident or, at least, an incident. ICAO initially tried to develop a "radiotelephony speech" based on simplified English. This sort of speech is called "ICAO standard phraseology". It was developed to cover many circumstances, which might occur during flights. They include routine events, non-routine events and some predictable emergencies.

In 1998, the ICAO Assembly took note of several accidents and incidents where the language proficiency of pilots and air traffic controllers was a causal or contributory factor. The worst accident in aviation history, in terms of fatalities, which occurred at Tenerife on March 27, 1977 was one of those accidents of which miscommunication in English was cited as a contributory factor. ICAO also discovered from linguistic research that there is no form of speech more suitable for human communication than natural language (ICAO, 2004). It means that "simplified English" or radiotelephony alone is not enough for aviation communication.

The ICAO provisions concerning standardized English language testing requirements and procedures were developed and the Proficiency Requirements in Common English Study Group (PRICESG) was established in 2000 to assist ICAO in the following aspects:

a) To carry out a comprehensive review of existing provisions concerning all aspects of air-ground and ground-ground communications in international civil aviation, aiming at the identification of deficiencies and/or shortcomings;

b) To develop ICAO provisions concerning standardized English language testing requirements and procedures; and

c) To develop minimal skill level requirements in common usage of the English language.

In March 2003, ICAO adopted amendments to its annex relating to language proficiency in international civil aviation. These amendments stipulate that pilots and air traffic controllers be required to demonstrate a certain level of English language proficiency in the use of both ICAO standard phraseology and plain language by March 2008.

In 2004, ICAO issued the “Manual on the Implementation of ICAO Language Proficiency Requirements” to be used as a guidance for those affected by these requirements. This manual mentions about the proficiency level requirements, the rating scales and the guidance in selection and/or development of suitable and effective language tests. The organization did not produce its own test to be utilized for this purpose. It just established the testing requirements and left the development of tests and test procedures to states, airlines and training organizations with the state aviation authority maintaining oversight responsibility.

The outcome from the PRICESG was the amendment of the ICAO Standards and Recommended Practices (SARPs) relating to language use in aeronautical radiotelephony communications. It requires flight crew and air traffic controllers to demonstrate language proficiency used in aeronautical communication.

Testing language proficiency is nothing new in the world of language assessment. There are many English language proficiency tests for both general purposes and specific purposes. *Test of English as a Foreign Language (TOEFL)* and *Test of English for International Communication (TOEIC)* are examples of those well-known international standardized tests. In theory, all tests are developed for some purposes. The English language proficiency test for pilots is intended to evaluate how an individual pilot is able to actually use English language appropriately in an aviation context. In such a context, pilots and air traffic controllers communicate verbally. This means speaking and listening skills are required. Furthermore, in order to avoid ambiguity under normal situations, they normally use standard phraseologies that are established by ICAO. However, when circumstances differ, pilots and air traffic controllers are expected to be able to use plain English as clear and as concise as possible. Therefore, this kind of test leads itself to an aviation context for testing. Indirect tests of grammar, reading, or writing are inappropriate. Other kinds of tests such as English for Academic Purposes (EAP) tests or other English for Occupational Purposes (EOP) tests like English for Business tests are also inappropriate. Therefore, those well-known tests cannot be used for this particular purpose. They may be used only for pre-training assessment or for screening.

This kind of ICAO language proficiency required test is considered as a very high stakes test. The safety of airline passengers depends on the effectiveness of pilot and air traffic controller communications. The test results will also have an impact on the career of pilots and controllers tremendously. The role of raters in this case is thus very crucial. They should not only be able to rate the test takers‟ language proficiency but also to identify deficiencies in the test takers‟ performance concerning ICAO‟s six criteria i.e. pronunciation, structure, vocabulary, fluency, comprehension, and interactions. They should also be able guide them towards appropriate language learning activities so that the pilots and controllers can focus their efforts to improve their language proficiency and language test performance (ICAO, 2004).

Since ICAO clearly states in its manual that “Direct, communicative proficiency tests of speaking and listening abilities are appropriate assessment tools for the aviation industry …” (ICAO, 2004:6-8), this kind of test requires qualified raters to perform the rating. To help ensure a comprehensive evaluation of each test-taker, ICAO requires at least two raters to be used to reduce the possibility of rater error. It also requires that at least one of them is a language expert (ICAO, 2008:22). Therefore, ICAO proposes two

kinds of raters, namely linguistic raters and operational raters. These two kinds of raters differ in their background knowledge. Linguistic raters are those who have linguistic knowledge and their assessment will focus on linguistic features of a test taker‟s performance while operational raters are those who have working knowledge of professional standards and procedures of radiotelephony communications and their assessment will focus on the appropriateness of a test taker‟s performance (ICAO, 2004).

In view of the fact that raters are considered as a critical factor in assessing pilot speaking proficiency and there is currently no definite conclusion concerning the use of raters. The issue of employing two kinds of raters is still debatable. Is it necessary to use two raters which means more time and cost consuming? Is it possible to use just one kind of raters to save both time and cost in rating? If so, do those linguists, alone, truly understand the type of aviation English used by pilots and air traffic controllers and grant the scores accordingly? On the other hand, are those experts in the field of aviation, exclusively, able to accurately identify the strengths and weaknesses in the test-takers‟ performance as required by ICAO and, of course, eventually able to rate the test-takers‟ English proficiency as it should be? Even if both kinds of raters end up with the same score for a particular test-taker, do they award that same score because of the same reasons? This study focuses on the important aspects that may affect their ratings including their educational and professional backgrounds. It also concentrates on the rating in the context of Thai pilots who are non-native speakers of English.

1.2 Research questions

Specifically, this study attempted to answer the following four research questions:

1.2.1 Does the different background knowledge of raters have any effect on their ratings of Thai pilot speaking ability?

1.2.2 Does rater training have any effect on their ratings of Thai pilot speaking ability?

1.2.3 Do the different background knowledge of raters and their training have any interactive effects on their ratings of Thai pilot speaking ability?

1.2.4 What are other factors affecting the decision making of raters in rating Thai pilot English speaking proficiency?

1.3 Research Objectives

The purposes of this study are:

1.3.1 To investigate the effects of the different background knowledge of raters on their ratings of Thai pilot speaking ability.

1.3.2 To explore the effects of rater training on their ratings of Thai pilot speaking ability.

1.3.3 To examine the interactive effects between rater background knowledge and their training with their ratings of Thai pilot speaking ability.

1.3.4 To examine other factors affecting the decision making of raters in their ratings of Thai pilot speaking ability.

1.4 Statement of hypotheses

The hypotheses concerning the relationship between rater background knowledge and rater training on rating Thai pilot English speaking proficiency are:

$H^1$1: The linguistic raters will rate test takers‟ performance significantly and differently from operational raters ($p \leq .05$).

$H^1$2: The raters who are trained in any rater training course will rate significantly and differently from those who are not trained ($p \leq .05$).

$H^1$3: There are significant effects among types of raters, rater training and rating performance ($p \leq .05$).

1.5 Scope of the study

This study focuses on two types of the background of raters, namely linguistic and operational raters in terms of their English linguistics and operation knowledge only. It does not concern any other kinds of personal background such as age, gender, etc. The study was administered with 10 linguistic raters and 10 operational raters.

The test, which is the data source, used in this study is called RELTA, which stands for “RMIT English Language Test for Aviation”. RELTA is a standardized test

developed by RMIT (Royal Melbourne Institute of Technology) English Worldwide, a global English language learning institution based in Melbourne, Australia. It is a part of RMIT Training Pty Ltd, a wholly owned commercial subsidiary of RMIT University, which is one of Australia‟s largest universities. The test was an early version of RELTA that was conducted with Thai pilots working for Thai Airways International PLC. Three randomly selected of these RELTA speech samples from three different proficiency levels conducted with those Thai pilots were used.

1.6 Limitations of the study

The researcher focused only on finding the effects of rater background knowledge, rater training and other factors affecting the decision-making of raters with different background knowledge and rater training in their rating of pilot English speaking proficiency.

In addition, the participants in this study included 10 operational raters who were Thai pilots from Thai Airways International PLC only. Operational raters from other airlines such as Bangkok Airways, Thai Air Asia, etc. or operational raters from other fields such as air traffic controllers were not included in this study. Therefore, the scores from operational raters were based solely on these 10 Thai Airways International pilots. The results of the study may not be applied to the operational raters from other agencies. Other raters were linguistic raters who are English language teachers. Four of them were from Thai Airways International Flight Crew Language Training Department while the other one was from the Civil Aviation Training Institute. The other five were English language teachers from various institutions. Speech samples from three levels of RELTA (Level 3, Level 4 and Level 5) were employed. The other levels (Level 1, Level 2 and Level 6) were excluded because their English proficiency levels were so obviously different that they could easily be distinguished. The scoring was based on the criteria set by ICAO and RMIT.

The selection of the factors affecting raters‟ decision-making was based on some previous research findings in speaking assessment. They are raters‟ educational and rating background; raters‟ mental conditions; raters‟ physical conditions; physical settings; raters‟ rating strategies; test tasks and speech samples; interviewer/interlocutor effects;

candidates/test-takers; rating scales and descriptors; the cut-off score; personal relationship between raters and candidates; scoring techniques; and raters‟ harshness/leniency.

1.7 Assumptions of the study

1.7.1 All raters honestly did the ratings with their best effort.

1.7.2 The operational raters are experienced line pilots, not newly recruited pilots. They are familiar with the working knowledge of professional standards and procedures of radiotelephony communications.

1.7.3 The effect of the rating setting on the test scores was kept to minimum since the rating setting was arranged in the similar environment and the ratings were administered individually in an isolated room.

1.8 Definition of terms

1.8.1 Thai Airways International Public Company Limited (THAI) is a Thai government enterprise conducting airline business and Thailand‟s national flag carrier.

1.8.2 THAI Pilots are pilots working as employees for Thai Airways International Public Company Limited.

1.8.3 International Civil Aviation Organization (ICAO) is an agency under the United Nations (UN) responsible for international aviation safety. Its headquarters is located in Montreal, Canada.

1.8.4 RELTA is the English for Occupational Purposes (EOP) test of speaking proficiency in English for pilots. It was developed by RMIT English Worldwide, which is a global English language learning institution - a part of RMIT Training Pty Ltd, a wholly owned commercial subsidiary of RMIT (Royal Melbourne Institute of Technology) University, which is one of Australia‟s largest universities, based in Melbourne, Australia.

1.8.5 Linguistic raters are raters who have background in linguistics and/or English language teaching.

1.8.6 Operational raters are raters who have working knowledge of professional standards and procedures of radiotelephony communications e.g. pilots or air traffic controllers.

1.8.7 Trained raters were raters who have passed either TOEIC language proficiency interviewer/rater training course in 2006 or TRAINAIR Standardized Training Package (STPs) interviewer/rater training course organized by Civil Aviation Training Center (CATC) in 2009, or both.

1.8.8 Untrained raters were raters who have not passed either TOEIC language proficiency interviewing/rating training course in 2006 or TRAINAIR Standardized Training Package (STPs) interviewer/rater training course organized by Civil Aviation Training Center (CATC) in 2009, or both.

1.9 Significance of the study

The research results would be advantageous in the following two main aspects:

1.9.1 Theoretical contribution

This kind of ICAO required language proficiency test is considered a very high stakes test. The safety of airline passengers depends on the effectiveness of pilot and air traffic controller communications. Furthermore, the outcome of the test will impact on the career of pilots and controllers since ICAO requires pilots, who operate flights internationally, to acquire the minimum Level 4 language proficiency. This research study was one of the very first of its kind in the field of aviation English language proficiency assessment in Thailand. The results would reflect some theoretical aspects of the controversial and debatable issue of utilizing different kinds of raters in this high stakes assessment. The results of the study would also provide some insights about the issues of rater‟s background knowledge and rater training in this kind of EOP assessment.

In addition, information on different sources related to rating aviation English speaking proficiency would also be obtained.

1.9.2 Practical contribution

People in related fields such as aviation regulatory bodies e.g. the Thai Department of Civil Aviation, test administrators, test providers, test takers, stakeholders and other interested persons could benefit from the research findings as follows:

- The results of the study would provide suggestions for selecting raters used in English language proficiency test for pilots.

- It would also provide suggestions on suitable selection and administration of raters in accordance with ICAO requirements in an English proficiency test for pilots in Thailand.

- Some factors such as raters‟ educational and professional backgrounds that affect the rating would be obtained and would be useful in the future rating.

- People who are interested in speaking assessment could use the research results to conduct further studies.

CHAPTER II
REVIEW OF THE LITERATURE

2.1 Introduction

This study investigates the effects of the raters‟ background and their training in the ICAO required assessment of Thai pilots‟ English speaking proficiency. The differences between the test results from the ratings of two types of raters, namely linguistic raters and operational raters were also studied.

The findings helped provide some insights about the issues of raters‟ background and rater training in EOP assessment and provide suggestions for further studies concerning raters‟ performance with different backgrounds.

This chapter sets out a review of related literature beginning with the definitions of English for Occupational Purposes (EOP) as a branch of ESP (English for Specific Purposes) and testing language for specific purposes followed by the characteristics of English used in aviation contexts or Aviation English, both ICAO standard phraseology and plain English. Next, it explores the history of oral proficiency tests. Then, the definitions of raters and rating scales are explained. An overview of ICAO rating scales and ICAO aviation language testing requirements are specified. Finally, the relationship between raters, ratings, rater training and the terms „inter-rater reliability‟ and „intra-rater reliability‟ are described.

2.2 English for Occupational Purposes (EOP)

English for Occupational Purposes (EOP) is one of the two branches of ESP (English for Specific Purposes) differentiated according to whether the learner requires English for work or for study. The other branch is English for Academic Purposes (EAP) which the learner requires English for academic study. Hutchinson and Waters (1987) proposed the concept of English Language Teaching (ELT) in the form of a tree of ELT (as shown in Figure 1), which represents some of the common divisions that are made in ELT.

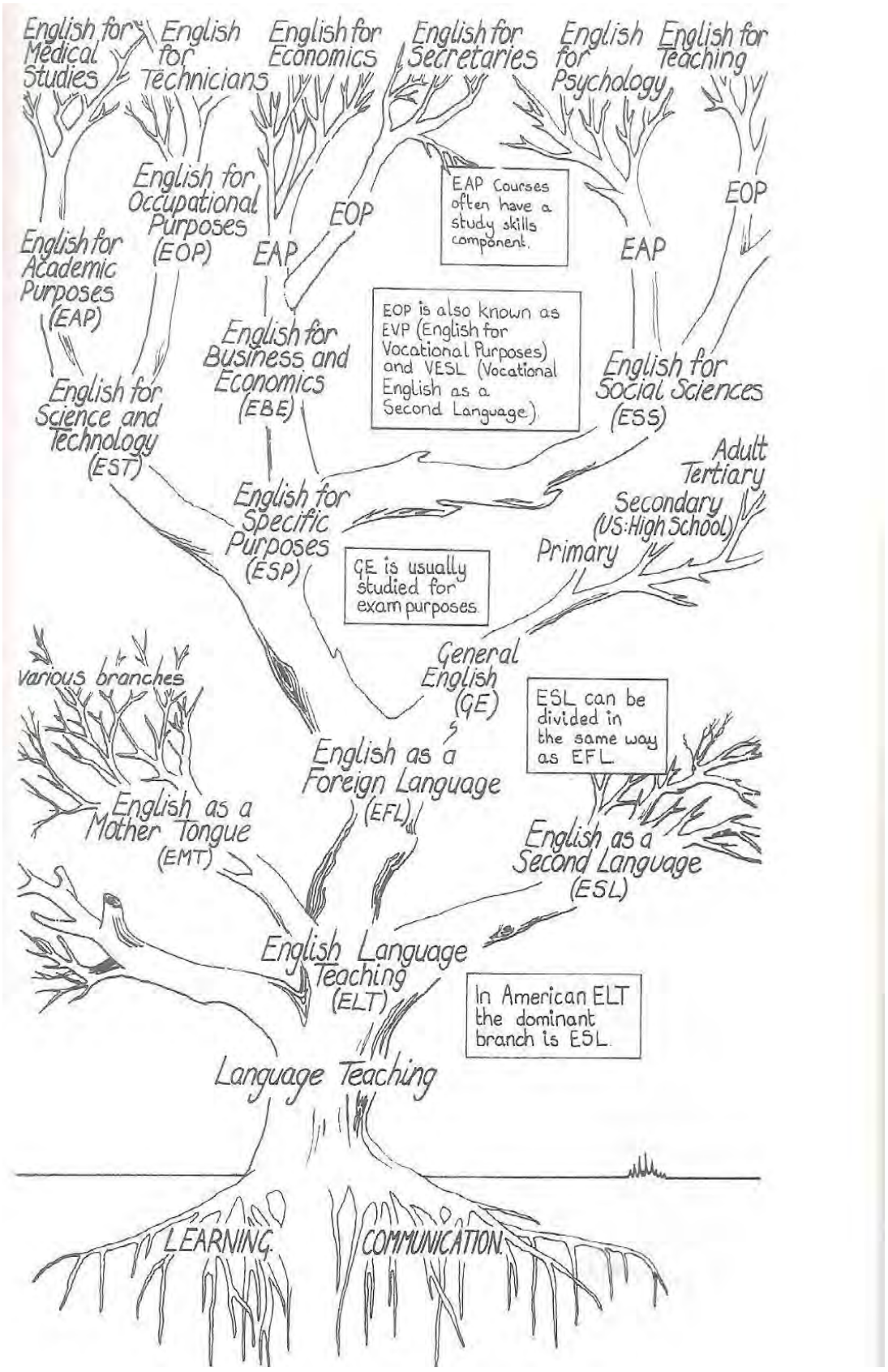

Figure 2.1: Tree of ELT (Hutchinson & Waters, 1987: 17)

Robinson interestingly defines ESP as “a type of ELT…it is goal-oriented, students study ESP not because they are interested in the English language as such but because they have to perform a task in English” (Robinson, 1989 cited in Davies, 2001:136). In case of EOP, students study English because they have to perform their jobs or their work in English.

In the perspective of Orr (2002:1), ESP is a subset of the English language that is required to carry out specific tasks for specific purposes. It is also “a branch of language education that studies and teaches subsets of English to assist learners in successfully carrying out specific tasks for specific purposes”.

Dudley-Evans (1998:5) explains that “ESP has traditionally been divided into two main areas: English for Academic Purposes (EAP) and English for Occupational Purposes (EOP)”. The examples of EAP are English for Science and Technology (EST), English for Medical Purposes (EMP), and English for Legal Purposes (ELP). These EAP studies are investigated for academic purposes, which differ from those that are learnt for occupational purposes even though they may be in the same discipline. For example, English for (academic) Medical Purposes is designed for medical students while English for (occupational) Medical Purposes is for practicing doctors.

There are some unique characteristics in English for Specific Purposes which differentiate one discipline from the others and, sometimes, from General English. The work of register analysis by Swales (1988) reveals that certain grammatical and lexical forms in Scientific and Technical English are used much more frequently. For example, present simple tense is the predominant tense and the passive voice is used much more frequently than in General English. Some semi- or sub-technical vocabulary e.g. „consists of“, „contains“, „enables“ are presented more in scientific and technical writing than in general contexts. This kind of register of using words or phrases in a particular way also happens in English for Occupational Purposes such as in English for Legal Purposes (legal language) and English for Aviation Purposes (aviation English) (see more details of aviation English in 2.3 below).

When testing gets involved with ESP, it turns out to be Testing Language for Specific Purposes (LSP). Douglas (2000:1) defines Testing Language for Specific

Purposes (LSP) as “that branch of language testing in which the test content and test methods are derived from an analysis of a specific language use situation, such as Spanish for Business, Japanese for Tour Guides, Italian for Language Teachers, or English for Air Traffic Controllers”. The two aspects of LSP testing which distinguish it from general purpose language testing are the authenticity of tasks and the interaction between language knowledge and specific purpose content knowledge. Douglas (2001:40) also defines specific purpose language ability in LSP testing as “specific language ability results from the interaction between specific purpose background knowledge and language ability, by means of strategic competence engaged by specific purpose input in the form of test method characteristics”. From this definition, the English language proficiency of pilots and air traffic controllers must be brought about by the interaction between their background knowledge in aviation and their English language ability.

In a specific purpose test development, LSP testing requires an analysis of a target language use situation and the cooperation between language testing specialists and experts in the field in constructing LSP tests. The material the test is based on must also engage test takers in a task in which both language ability and knowledge of the field interact with the test content in a way which is similar to the target language situation (Douglas, 2000). Another characteristic of LSP testing is the use of technical language that people who work in the field must be able to control.

2.3 Aviation English

In order to understand pilots‟ use of English, the term that must be mentioned and clarified is „aviation English‟.

Aviation English is defined in the Manual on the Implementation of ICAO Language Proficiency Requirements as “a comprehensive but specialized subset of English broadly to aviation, including the plain language used for radiotelephony communications when phraseologies do not suffice” (ICAO, 2004:4-8).

It is also defined in the same manual that Radiotelephony English is “a sub-category of aviation English” (ICAO, 2004:4-8) It is the language used in radiotelephony communications. Radiotelephony English “includes, but must not be limited to, ICAO

phraseology and can require the use of general English at times" (ICAO, ibid) and ICAO phraseology is "the standardized words and phrases approved for radiotelephony communications by ICAO which have been developed over years and represent a very narrow, specialized and rigid subset of language" (ICAO, ibid).

The primary objective of using English for pilots is to communicate with air traffic controllers. This has to be done clearly, concisely and unambiguously. Only one misunderstanding may lead to a catastrophic disaster. However, not all pilots and air traffic controllers are English native-speakers nor fluent in English.

International Civil Aviation Organization recognizes this problem and the need for a standard and unambiguous language system which can be easily used by all concerned (ICAO, 2004). This language system is the English-based radiotelephony system as English is accepted as the "lingua franca" of aviation (Crystal, 1997: 98-99).

The English language of international air traffic control, which presents international safety, has standard terminology and phraseology to avoid ambiguity between pilots and air traffic controllers. Pilots do not talk in a normal way to air traffic controllers. They use a restricted vocabulary and a fixed set of sentence patterns, which aim to express unambiguously in all possible air situations. They use terms such as "Roger", "Wilco" and "Mayday"; phrases such as "Maintaining 3000" and "Runway in sight"; and the use of a phonetic alphabet to spell out code names or call signs e.g. "Alpha" for A, "Bravo" for B, etc.

Most of these aviation registers are recommended by the International Civil Aviation Organization (ICAO) and adopted by 194 contracting states worldwide (ICAO, 2004). Like the other specialized registers of occupational groups, they are developed initially from the desire for quick, efficient, and precise communication between people who share experience, knowledge and skills.

The obvious distinguishing feature of aviation English registers is the phraseology. Phrases like "Clear for takeoff", "Clear to land" to give or acknowledge air traffic control (ATC) clearances and "Climbing to Flight level 310", "Leaving 3500" to report actions being taken, are examples of phrases specific to aviation.

2.4 Characteristics of Aviation English

ICAO Alphabet spellings

When proper names, service abbreviations and words of which the spelling is doubtful are spelled out in aviation radiotelephony, the following alphabet spellings are used (International Civil Aviation Organization Annex 10, 2001:5-4):

Letter	Alphabet representation
A	AL FAH
B	BRAH VOH
C	CHAR LEE
D	DELL TAH
E	ECK OH
F	FOKS TROT
G	GOLF
H	HO TELL
I	IN DEE AH
J	JEW LEE ETT
K	KEY LOH
L	LEE MAH
M	MIKE
N	NO VEM BER
O	OSS CAH
P	PAH PAH
Q	KEH BECK
R	ROW ME OH
S	SEE AIR RAH
T	TANG GO
U	YOU NEE FORM
V	VIK TAH
W	WISS KEY
X	ECKS RAY
Y	YANG KEY
Z	ZOO LOO

For example, the name “Chula” is spelled out as “CHAR LEE, HO TELL, YOU NEE FORM, LEE MAH, AL FAH”.

Alphabet spelling exceptions

However, not all abbreviations are under the rule of the alphabet spellings. Some of these registers are normally spelled and pronounced, for example:

VOR, which stands for “VHF (Very High Frequency) Omni Range”, is pronounced as “vee-o-ar”, not “VIK TAH, HO TELL, FOKS TROT”.

DME (Distance Measuring Equipment) is pronounced as “dee-em-ee”, not “DELL TAH, MIKE, ECK OH”. This term is applied as a unit of distance measurement, not the equipment itself. It is equal to the distance of one nautical mile (1.85 km) from a DME station. For instance, “Six DME” means that the aircraft is at six nautical miles from a DME station.

TCAS (Traffic alert and Collision Avoidance System) is pronounced as “tee-cas”, not “TANG GO, CHAR LEE, AL FAH, SEE AIR RAH”.

ACAS (Airborne Collision Avoidance System) is pronounced as “ay-cas”, not “AL FAH, CHAR LEE, AL FAH, SEE AIR RAH”.

PF (Pilot Flying = the pilot who is in control of the aircraft) is pronounced as “pee-ef”, not “PAH PAH, FOKS TROT”.

PNF (Pilot Not Flying = the pilot who assists PF, he is not in control of the aircraft) is pronounced “pee-en-ef”, not “PAH PAH, NO VEM BER, FOKS TROT”.

IMC (Instrument Meteorological Condition) is pronounced as “ai-em-see”, not “IN DEE AH, MIKE, CHAR LEE”.

VMC (Visual Meteorological Condition) is pronounced as “vee-em see”, not “VIK TOR, MIKE, CHAR LEE”.

CPDLC (Controller-Pilot Data Link Communication) is pronounced as “see-pee-dee-el-see”, not “CHAR LEE, PAH PAH, DELL TAH, LEE MAH, CHAR LEE”.

The reasons for these exceptions are that these abbreviations are so well known and extensively used that they give no ambiguity when spoken. On the other hand, if they are phonetically spelled alphabets, they may cause some misunderstanding.

Transmission of numbers

As stated in International Civil Aviation Organization Annex 10 (2001:5-5), some numbers were distinguishably pronounced to be certain that there would be no ambiguity. However, nowadays fewer and fewer pilots pronounce them in this manner.

Numeral or numeral element	Pronunciation
1	Wun
2	Too (not Two)
3	Tree (not Three)
4	Fow-er (not Fow)
5	Fife
6	Six
7	Sev-en
8	Ait
9	Nin-er (not Nine)
0	Ze-ro

Aiguo (2007) interestingly explains the reasons for these distinguished pronunciations. He clarifies that in English air communication „„3“ is read out /Tree/ instead of /Thri:/, „„4“ is read out /Fow-er/ and „„9“ is read out /Nin-er/ instead of /nain/ as usual. Since the sound /h/ is interdental and voiceless, it is difficult to be heard by the listener in communication, so the /h/ sound is replaced by alveolar and plosive /t/ in air communication, and the word „„thousand“ is pronounced as /Tou-Sand/. Therefore, /tr/ is likely to replace /hr/ in this case with the consideration of efficiency and clarity. The pronunciation of the number „„4“ gets easily confused with that of the preposition „„for“, so the vowel /er/ is added (/Fow-er/) to distinguish the two sounds. In pronouncing number „„9“, the second /n/ sound in /nain/ is a nasal and this makes it difficult to be

heard too, so it would be safer and easier to be heard if it is read as /Nin-er/, with a vowel /er/ added to it.

All numbers except whole hundreds, whole thousands and combinations of thousands and whole hundreds are transmitted by pronouncing each digit separately. Whole hundreds and whole thousands are transmitted by pronouncing each digit in the number of hundreds or thousands followed by the word “hundred” or “thousand” as appropriate. Combinations of thousands and whole hundreds are transmitted by pronouncing each digit in the number of thousands, followed by the word “thousand”, followed by the number of hundreds, followed by the word “hundred”, for example:

Number	Transmitted as
10	One Zero
75	Seven Five
583	Five Eight Three
600	Six Hundred
5000	Five Thousand
7600	Seven Thousand Six Hundred
11000	One One Thousand
18300	One Eight Thousand Three Hundred
38143	Three Eight One Four Three

Transmission of numbers in hundred and thousand exceptions

1) Numbers containing a decimal point are transmitted separately, even in whole hundred or whole thousand, for example:

Number:	Transmitted as:
100.3	One Zero Zero Decimal Three
1200.4	One Two Zero Zero Decimal Four
2000.5	Two Zero Zero Zero Decimal Five
38143.9	Three Eight One Four Three Decimal Nine
45000.1	Four Five Zero Zero Zero Decimal One

2) When transmitting time, only the minutes of the hour are normally required. Each digit is pronounced separately. However, the hour should be included when any possibility of confusion is likely, for example:

Time:	Transmitted as:
0900 (9:00 A.M.)	Zero Nine Zero Zero
1000 (10:00 A.M.)	One Zero Zero Zero
1643 (4:43 P.M.)	One Six Four Three or Four Four Three

Three- and Four-letter location indicators

ICAO has set up two systems to refer to every airport in the world. So, there is no need to write or print the whole name of such airports which may be too long, or to eliminate any uncertainty or misunderstanding between air traffic controllers and pilots. They are called 3-letter and 4-letter location indicators. One airport has both three- and four-letter location indicators, such as Bangkok International Airport three-letter location indicator is BKK and four-letter is VTBD, for example:

ATC: “Request your destination.”

(The air traffic controller wants to know where the pilot is going to)

Pilot: “Echo Sierra Sierra Alpha”

(The pilot replies that he is going to Stockholm/Arlanda International Airport in Stockholm, Sweden).

From the example above, if the pilot does not use the 4-letter location indicator, he may have to speak longer. Moreover, if the air traffic controller is not familiar with or has never heard the name of the airport, he may not be able to make a good guess at all. By using this kind of register, both parties can thoroughly understand each other.

Instead of spelling these location indicators separately, some of them can be read like a word. This is not official but broadly used among pilots, for example:

Pilot A: “Where are you heading?”

Pilot B: “KIX, then LAX”

(KIX, pronounced as “kigs”, stands for Osaka/Kansai International Airport in Osaka, Japan. LAX, pronounced as “lags”, stands for Los Angeles International Airport in Los Angeles, California, USA.)

These three-letter and four-letter location indicators are also useful in identifying the airport located in a city that has more than one airport in the vicinity e.g. LGW/EGKK is for London/Gatwick, LHR/EGLL for London/Heathrow and STN/EGSS for London/Stansted.

Words and phrases

The philosophy behind standardized words and phrases in aviation is the same as alphabet spelling and number transmission, which is to avoid ambiguity i.e. one word – one meaning. Most of these words and phrases are regulated by ICAO (International Civil Aviation Organization Annex 10, 2001:5-6, 5-7) for example:

Phrases	Meaning
“Acknowledge”	“Let me know that you have received and understood this message”
“Affirm”	“Yes”
“Break”	“I hereby indicate the separation between portions of the message”
“Break Break”	“I hereby indicate the separation between messages transmitted to different aircraft in a very busy environment”
“Charlie Charlie”	“Yes” (The same meaning as “Affirm” but this phrase is not recognized by ICAO)
“Confirm”	“Have I correctly received the following …?” or “Did you correctly receive this message?
“Correction”	“An error has been made in this transmission (or message indicated). The correct version is …”
“Mayday”	The aircraft is in a distress situation. It means that grave and imminent danger is present, and immediate assistance is requested.
“Negative”	“No” or “Permission not granted” or “This is not

	correct".
"Negative contact"	To acknowledge traffic information that "I cannot see the informed traffic".
"Pan Pan"	The aircraft is in urgency situation. It wishes to give notice of difficulties which compel it to land without requiring immediate assistance.
"Roger"	"I have received all of your transmission". (This is obsolete.)
"Squawk"	To instruct setting of transponder code on the aircraft transponder. One set comprises four digits e.g. when an air traffic controller instructs a pilot to "squawk zero seven four four", it means that he wants the pilot to set "0744" on the aircraft transponder.
"Squawk Ident"	To instruct pilot to depress the identification button on the aircraft transponder.
"Traffic in sight"	To acknowledge traffic information that "I can see the informed traffic".
"Wilco"	"I understand your message and will comply with it". (This is also obsolete.)

Routines and formulas

An interesting feature of aviation communications is the use of routines to begin the conversations. It is the same as sports commentaries in England, livestock auctions in New Zealand, tobacco auctions in the United States or North American ice hockey commentaries. These registers are all characterized by the extensive use of oral formulas. The formulas involve a small number of fixed syntactic patterns and a narrow range of lexical items.

The following excerpt comes from an aviation communication between a Thai Airways International pilot and a Bangkok air traffic controller (ATC). The pilot is requesting a clearance to fly to London/Heathrow International Airport, for example:

Pilot: "Bangkok control. (This is) THAI 910. Request ATC clearance to London/Heathrow. Flight Level 310."

ATC: "(Bangkok) clears THAI 910 to London/Heathrow (Airport) (via) Alpha 1. Flight planned route. Flight Level 310. Frank 1 departure. Limla transition. Squawk 0721"

Pilot: "THAI 910 is cleared to London/Heathrow. Alpha 1. Flight planned route. Flight Level 310. Frank 1 departure. Limla transition. Squawk 0721."

ATC: "THAI 910. (Your) read back is correct. Contact Ground 121.9"

Pilot: "THAI 910"

This example is composed of a set of pre-determined formulas which are the "initial contact" formula, the "request" formula, the "ATC clearance issuance" formula, the "read back" formula and the "acknowledgment" formula. They can be described as follows:

(a) The initial contact formula

"Bangkok Control" = the addressee

"(This is) THAI 910" = the addresser

(b) The request formula

"Request …"

(c) The ATC clearance issuance formula

"THAI 910 …" refers to the addressee;

"is cleared to …" refers to the destination;

"Alpha 1" refers to the name of the airway;

"Flight planned route" refers to the route which was filed;

"Flight Level 310" refers to the altitude to which is cleared to climb and maintain;

"Frank 1 departure" refers to the name of SID (Standard Instrument Departure);

"Limla transition" refers to the name of the transition from SID to airway

"Squawk 0721" refers to the set of transponder code to be selected.

(d) The read back formula

"THAI 910 is cleared to …" is the read back. To make sure that there is no ambiguity or miscommunication.

(e) The acknowledgment formula

"THAI 910" is the acknowledgment. To accept any message, the pilot must end the conversation by stating his call sign.

This register is characterized by a very restricted range of lexical and syntactic variation. Moreover, the specific features of the formulas are not arbitrary, but motivated by the demand of the context. Finally, the sound patterns of aviation communication are also distinctive. They must be as slow and clear as practicable, no matter how fluent the speakers are in English, to avoid possible confusion by those persons using a language other than the one of their national languages.

Robertson (1988) categorizes the normal patterns of a flight as pilots actually perform during their line of duties into four parts. Each part is divided into sections which follow the normal sequence of events for each phase of flight. Most of them consist of requests and permissions and/or instructions. After receiving the answers, pilots must acknowledge the replies from air traffic controllers in order to confirm that they really have received and thoroughly understand them.

The flight patterns are as follows:

Part 1 - Pre-flight to line-up

1.1 Departure information

1.1.1 Departure information (routine)

This section comprises the pilot‟s request for the departure airport weather information e.g. the runway in use, surface wind, temperature, etc. and the reply from the air traffic controller. The pilot may write this information on a piece of paper.

1.1.2 Departure information (ATIS – Automatic Terminal Information Service)

This section is the automatic transmission of the recorded information which is updated regularly i.e. every half an hour.

1.2 Route clearances

This section consists of the pilot‟s request for the route to be flown and the details of the route with the permission to fly that route from the air traffic controller. The pilot writes these clearances on his/her copied flight plan in order to confirm that it complies with his/her filed flight plan.

1.3 Start-up

In this section, the pilot requests permission to start the engine(s) and the ground controller approves that request.

1.4 Push-back

The pilot asks for the consent to be pushed out of his/her parking position and the ground controller gives the approval for that.

1.5 Taxiing

In this section, the pilot requests permission to taxi to the runway in use. The ground controller grants that and provides the direction to taxi to the runway.

1.6 Line-up

The air traffic controller permits the pilot to line up the aircraft on the runway in use.

Part 2 - Take off to top of climb

2.1 Take-off

The air traffic controller gives the permission to take off to the pilot.

2.2 Initial climb

The air traffic controller gives initial clearances to the pilot.

2.3 Climb

The air traffic controller gives further clearances to the pilot to continue climbing to the top-of-climb.

2.4 Top-of-climb

The pilot reports to the air traffic controller that he/she has reached the cruising altitude.

Part 3 - Cruise to descent

3.1 Volmets

The pilot receives the recorded weather broadcasts of the relevant airports i.e. the destination airport and the alternate.

3.2 En route: Position reports

The pilot reports his/her position to the air traffic controller at each specific point.

3.3 En route: Climb

The pilot asks for permission to climb to higher altitude as his/her aircraft weight has decreased because of the fuel used.

3.4 En route: Traffic information

The air traffic controller informs the pilot about the other aircraft in the vicinity of his/her aircraft.

3.5 Descent

The pilot requests permission to leave his/her cruising altitude in order to land at the destination airport. The air traffic controller assigns the new altitude to the pilot.

Part 4 - Approach to landing

4.1 Arrival: ATIS

The pilot receives the ATIS which has a similar pattern to the one during the pre-flight phase except this is for landing runway.

4.2 Approach

The pilot contacts the air traffic controller in order to get clearance for the type of approach e.g. ILS, VOR, etc.

4.3 Final approach and Landing

The air traffic controller gives final approval for the approach and permission to land. The pilot must acknowledge these clearances.

4.4 After landing

The pilot reports when his/her aircraft is clear of the landing runway. The ground controller gives him/her the taxi instructions to the parking position.

This radiotelephony may be considered in terms of skills and language functions. Most of these radiotelephony utterances have functions of requesting and accepting information and require the skills of speaking and listening.

The purpose of using these phraseologies is to promote clarity and brevity. Still, it is widely acknowledged by operational and linguistic experts that no set of standardized phraseologies can fully describe all possible circumstances and responses (ICAO, 2004).

2.5 History of Oral Proficiency Tests

The term “oral test” appears in language testing prior to the Second World War. Still, by that time, it did not mean that test takers were required to „really“ speak in the test. Rather, it referred to the testing of pronunciation, usually required the test takers to write down the pronunciation of a written word using phonetic scripts. A speaking test was abandoned because of reliability problems. Because of that, language testing practitioners tried to concentrate on the „new-type“ multiple choice tests as reliable, objective measures of language ability (Fulcher, 2003: 2). This reflects the concern and the importance of the „reliability“ in language speaking tests since the early time of this kind of testing.

The first true speaking test used in North America was the College Board‟s English Competence Examination, introduced in 1930 for overseas students applying to study at US colleges and universities (College Entrance Examination Board, 1929, cited in Fulcher, 2003:2). The format of the speaking test is a conversation with ten topics prepared for the examiner. The criteria for assessment were fluency, responsiveness, rapidity, articulation, enunciation, command of construction, use of connectives, and

vocabulary and idioms. The examiner graded each examinee on the three-point scale of proficient, satisfactory, and unsatisfactory.

During the Second World War, the Army Specialized Training Program (ASTP) was established in 1942 to address the communication problems of American service personnel through the delivery of language programs that focused on speaking. However, the US government suspended the ASTP in 1944. After this, the Foreign Service Institute (FSI) was set up in order to teach foreign languages for American military personnel in overseas posts. In 1956, the FSI was given the responsibility to provide evidence of foreign language proficiency.

Nowadays, when referring to the oral proficiency test, the most well known test of its kind is the Oral Proficiency Interview (OPI). It is a structured procedure for the assessment of functional speaking ability and was developed through work initiated by the Foreign Service Institute (FSI) of the U.S. government in 1958 and the subsequent contributions of The Peace Corps, Educational Testing Service and the cooperative efforts of academic institutions from around the United States. It claims to assure reliability in assessing oral proficiency and it measures patterns of strengths and weaknesses, establishing a speaker's level of consistent functional ability as well as the clear upper limitations of that ability. It is administered face-to-face with two certified raters lasting from 30 minutes to an hour.

OPI assesses the candidate‟s listening comprehension and speaking proficiency and takes into consideration factors such as fluency, grammar, pronunciation, vocabulary, and ability to successfully work through various linguistic tasks. It consists of four stages; a warm-up, to include autobiographical information; level checks, to assess ability to perform linguistic tasks at a base level; level probes, to determine ability to perform linguistic tasks at the next higher base level; and a wind down, to put the candidate at ease.

The interview is rated on the U.S. ILR (Inter-agency Language Roundtable) 11-point scale of proficiency, from 0, no functional proficiency, to 5, educated native-speaker proficiency, with plus levels (0+,1+2+,3+,4+) assigned to those who demonstrate inconsistent proficiency at the next higher level.

The ACTFL (American Council on the Teaching of Foreign Languages) Oral Proficiency Interview was developed to evaluate speaking proficiency in a foreign language. It is a criterion-referenced, direct, face-to-face interview with only one interviewer present. The interview consists of five stages: the warm-up, level checks, probes, role-play, and wind-down. The role of the 'warm-up' is to put the interviewee at ease, to familiarize him/her with the pronunciation and way of speaking of the interviewer, and to generate topics which can be explored later in the interview. The 'level checks' allow the interviewee to demonstrate his/her ability to manipulate tasks and contexts at a particular level. If the interviewer is satisfied with the candidate's sustained performance, an attempt will be made to discover the 'ceiling', i.e. to elicit response at the higher level. 'Probes', thus, makes the candidate reveal a pattern of weaknesses. A 'role-play' serves as an additional check, to help the interviewer confirm the candidate's level. The 'wind-down' brings the interviewer down to a level comfortable for the candidate to end the OPI on a positive note. The entire interview lasts about 15 minutes in the case of a novice, and can be as long as 35 minutes if a series of probes and level checks are necessary. The interview is taped and a decision is made if the interviewer and a second rater agree on the level. In case of disagreement, the tape is sent to a third rater.

In the early l980s, ACTFL OPI proficiency scales developed out of the FSI (Foreign Service Institute) levels of oral proficiency. The American Council on the Teaching of Foreign Languages (ACTFL), the Educational Testing Service (ETS), and the ILR (Interagency Language Roundtable) began working on an adaptation of the OPI proficiency scale to be used in secondary schools and colleges. The result of that collaboration, the ACTFL Provisional Proficiency Guidelines, was published in l982. These guidelines made a number of changes in the OPI scale, yet were designed to be commensurate with it. First, the numerical designations of points on the scale were replaced with names that represent each level. Second, a further subdivision was made within the two lowest levels on the scale. Thus, Level 0 was renamed Novice and subdivided into Novice Low, Novice Mid, and Novice High, while Level 1 was renamed Intermediate and subdivided into Intermediate Low, Intermediate Mid, and Intermediate High. Level 2 was renamed Advanced, and Levels 3, 4, and 5 on the OPI scale were combined into a single level called Superior, because data had shown that few university graduates reach even Level 3. Following their publication, the Guidelines were widely distributed for comments throughout the foreign language teaching profession. Several hundred individuals were later trained to administer a face-to-face speaking test to assign

one of the proficiency levels defined in the Guidelines to each person tested. Because of their field-testing, the guidelines were determined to be an appropriate scale for assessing language proficiency among secondary and college-level students of foreign languages. Thus, following minor revisions, the word Provisional was removed, and the scale was republished in l986 as the ACTFL Proficiency Guidelines (ACTFL, 1999).

2.6 Raters

The dictionary of language testing defines a rater as *"the judge or observer who operates a rating scale in the measurement of oral and written proficiency."* (Davies et al., 1999:161) By this definition, it implies that a rater is a human, not an electronic rater that uses a computer to do the proficiency rating that is not covered in the scope of this study.

Richards and Schmidt (2002:441) define a rater as *"a person who assigns a score or rating to a test taker"s oral or written performance on the basis of a set of rating criteria."*

ICAO defines in the Manual on the Implementation of ICAO Language Proficiency Requirements that a rater is a suitably qualified and trained person who assigns a rating to a test taker"s performance in a test based on a judgment usually involving the matching of features of the performance to the descriptors on a rating scale (ICAO, 2004).

ICAO also classifies two types of raters as:

1. Linguistic rater – A rater whose assessment will focus on the linguistic features of a test taker"s performance in a test, and;

2. Operational rater – A rater with working knowledge of professional standards and procedures of radiotelephony communications whose assessment will focus on a test taker"s performance with regard to the holistic descriptors.

Valdes (2006), who was a member of ICAO Proficiency Requirements in Common English Study Group (PRICESG), additionally proposes the tentative requirements for ICAO language proficiency raters in an ICAO Proficiency Requirements in Common English Study Group (PRICESG) meeting as:

1. At least seven years experience of working as air traffic controllers or pilots or five years experience teaching English as a second language.

2. The level of the English language proficiency (speaking, listening, reading and writing is "proficient" or above), proved by the certification as follows:

- The IELTS examination – Academic version (the average score of 8.0 including 8.0 on speaking and listening accordingly; or

- The IELTS examination – General Training version (the average score of 8.0 including 8.0 on speaking and listening accordingly; or

- The Cambridge CPE examination (results A or B); and/or

- Level 6 (Expert) of language proficiency in accordance with ICAO language proficiency rating scale.

2.7 Rating scale

Another term, which is closely related to raters, is the "rating scale" or "proficiency scale". Rating scale is described as *"a technique for measuring language proficiency in which aspects of a person"s language use are judged using scales that go from worst to best performance in a number of steps."* (Richards and Schmidt, 2002:441). While Davies et al. (1999:153-4) explain the term equivalently as "proficiency scale" as *"a scale for the description of language proficiency consisting of a series of constructed level against which a language learner"sperformance is judged. Like a test, a proficiency (rating) scale provides an operational definition of a linguistic construct such as proficiency. Typically such scales range from zero mastery through to an end-point representing the well-educated native speaker"*.

Rating scales are important in tests of speaking because they are operationalizations of the construct that the test is supposed to measure (Fulcher, 2003:113). Moreover, the band descriptor, which is a statement describing the level of performance required of candidates at each point on a proficiency scale (Davies et al.,1999:43) is a major part of the „meaning" of the score, and delimits the type of inferences that can be made from the test score by the score user (Fulcher, 2003:113).

The importance of a rating scale can be realized by the quotation proposed by Lumley (2002:263). He stated, "In performance assessment, which relies on rating, there is an assumption that if a rating scale is developed in a valid way and raters are adequately trained to understand its content, then the scale will be used validly and reliably, and it will be possible to obtain good, or at least adequate, measurement".

Alderson (1991 cited in Fulcher 2003:89) suggests that there are three kinds of rating scales divided in terms of orientations, namely;

- User-oriented scales which are used to report information about typical or likely behaviors of a test taker at a given level,
- Assessor-oriented scales that are designed to guide the rating process, focusing on the quality of the performance expected,
- Constructor-oriented scales which are produced to help the test constructor select tasks for inclusion in the test.

A rating scale provides an operational definition of linguistic construct such as proficiency. These rating scales typically range from zero mastery through to an end-point representing the well-educated native speaker. One of the first most widely known of such scales should have been the Interagency Language Roundtable (ILR) rating scale which was built in 1958. The ILR scale was a set of descriptions of abilities to communicate in a language. It was originally developed by the United States Foreign Service Institute (FSI), the predecessor of the National Foreign Affairs Training Center (NFATC). Thus, it is also often called Foreign Service Levels. It consisted of descriptions of five levels of language proficiency. It was divided into five main categories, with „plus" levels (0+, 1+, 2+, 3+, 4+) assigned to those who demonstrate inconsistent proficiency at the next higher level. The ILR levels are: ILR Level 0 (No functional proficiency), ILR Level 1 (Elementary proficiency), ILR Level 2 (Limited working

proficiency), ILR Level 3 (Professional working proficiency), ILR Level 4 (Full professional proficiency), and ILR Level 5 (Native or Bilingual proficiency) (Fulcher, 2003; ILR, 2010a; ILR, 2010b). The details of these ILR levels are shown in Appendix A.

Another well-known rating scale is the ACTFL (American Council for the Teaching of Foreign Languages) scale. The American Council for the Teaching of Foreign Languages initially developed the scale in 1986. Therefore, it was called the ACTFL Proficiency Guidelines (1986). It was initially categorized as Novice, which is a non-survivor who relies on memorized materials and only reacts, does not initiate; Intermediate, which is a survivor who can create his/her own language, even if with many errors, can ask and answer questions and discuss daily events; Advanced, which is a person who has limited professional competence and can narrate, describe, and compare in any time frame also able to state opinions; and Superior, which is a person who has full professional competence and can go outside limited areas of competence and discuss a wide range of topics, also able to hypothesize and deal with abstract topics.

A significant change to the ACTFL Proficiency Guidelines (1986) was found in the division of the Advanced level into the High, Mid, and Low sublevels (ACTFL, 1999). This change reflects the growing need in both the academic and commercial communities to more finely delineate a speaker‟s progress through the Advanced level of proficiency. The new descriptors for Advanced Mid and Advanced Low are based on hundreds of Advanced-level language samples from OPI testing across a variety of languages. The presentation of these *Guidelines* was slightly different from previous versions. The full prose descriptions of each level (and, when applicable, its sub-levels) are preceded by clearly delineated „thumb-nail sketches‟ that are intended to alert the reader to the major features of the levels and to serve as a quick reference, but not in any way to replace the full picture presented in the descriptions themselves. Indeed, at the lower levels they refer to the mid rather than to the baseline proficiency, since they would otherwise describe a very limited profile and misrepresent the general expectations for the level.

The revision of the *ACTFL Proficiency Guidelines C Speaking* (ACTFL, 1999) is presented as an additional step toward more adequately describing speaking proficiency. This effort reflects a broad spectrum of experience in characterizing speakers‟ abilities

and includes a wide range of insights as a result of on-going discussions and research within the language teaching profession. These levels are classified as: Superior, Advanced-High, Advanced-Mid, Advanced-Low, Intermediate-High, Intermediate-Mid, Intermediate-Low, Novice-High, Novice-Mid, and Novice-Low (ACTFL, ibid). The details of the ACTFL levels are shown in Appendix B.

Many research indicate that the interpretation of rating scale by raters is one of the problems arising in this „subjective" assessment. Even though it is designed to help raters to make decision in their ratings, each individual rater has his/her own way of interpreting these scales and descriptors. A rating scale descriptor is "a statement which describes the level of performance required of candidates at each point on a proficiency scale" (Davies et al., 1999:43). In theory, raters refer to a rating scale in order to select a score to represent the candidate"s ability in the trait of interest (Upshur & Turner, 1999). In reality, each rater has a unique background that may affect his/her judgment (Brown, 1995; Elder, 1993). Interpretation of a rating scale is always an interest of many researchers. Lumley (1995) found differences in the interpretation of the rating scale used by trained ESL raters and medical practitioners. This finding confirmed Brown"s study about the perception of language-trained raters and experienced guides in 1995 that the two groups interpreted different criteria in different ways. Brown"s conclusion of her study is interesting. She remarked that "raters appear to have inbuilt perceptions of what is acceptable to them and these perceptions are formed to some extent by their previous experience" and "it appears that even the explicitness of the descriptors and the standardization that takes place in a training session cannot remove these differences" (Brown, 1995: 13). The possible implication of this remark is that if the descriptors are inexplicit, raters" perceptions are prone to base on their previous experience. Imprecise rating scales often results in holistic marking by raters (Weigle, 2002 cited in Knoch, 2009). That leads raters to use the overall or global impression of the candidates in their ratings instead of using an analytic rating scale, as it should be (Knoch, 2009). This misuse has a significant effect on this ICAO proficiency assessment since ICAO requires the lowest score in any criteria to be the overall score (ICAO, 2004). This study results demonstrate that all raters faced a degree of difficulties to explain how they interpreted the ICAO descriptors. Even though it is a common practice in language testing that the descriptors are categorized using adjectives like those mentioned in the ICAO descriptors (Knoch, 2009), each of the raters had dissimilar ideas of the descriptors i.e. „never",

„almost never", „rarely", „sometimes", „frequently" and „usually". This is one of the most commonly mentioned problems among raters. They thought that the descriptors were often too vague to arrive easily at a score (Knoch, 2009). This inexplicit interpretation of descriptors by each rater may affect his/her ratings to a certain extent.

2.8 ICAO Aviation Language Testing

Because the safety of airline passengers depends on the effectiveness of pilot and air traffic controller communications and the outcome of the test will affect the career of pilots and controllers, this language proficiency test is considered a very high stakes test. The ICAO language proficiency requirements point towards an aviation context for testing and requires proficiency tests of actual speaking and listening ability. In addition, the test should be work-related language proficiency test (ICAO, 2004). However, ICAO also emphasizes that ICAO standard phraselogies-only testing is not appropriate. The reason why ICAO requires two kinds of raters, namely linguistic and operational raters, is because, in terms of operational raters, ICAO clearly states that "the participation of operational experts, pilots and controllers or trainers in the rating process can add operational integrity to the process, as well as provide technical accuracy" (ICAO, 2004:6-4). In terms of linguistic raters, ICAO is concerned about candidates who do not pass the test "will want, and will deserve, accurate information about how their performance fell short of the target performance and in what areas they should focus their efforts to improve (their) performance" (ICAO: ibid.). Therefore, raters should not only be able to rate the test takers, they should also be able to identify deficiencies in the test takers" performance and guide them towards language learning activities so that they can focus their efforts to improve their language proficiency and language test performance later. This requires raters with background in linguistics or language teaching. This is the kind of information that linguists or language teachers can provide to candidates. ICAO concluded that the best practice in this kind of language proficiency assessment would call for at least two trained and calibrated raters, at least one of them is a language teacher (ICAO, 2004).

In conclusion, ICAO (2004: 6-5) emphasizes the critical characteristics of an appropriate testing system in the context of aviation language testing as follows:

1) It must be a proficiency test of speaking and listening;

2) It must be based on the ICAO Rating Scale and holistic descriptors;

3) It must test speaking and listening proficiency in a context appropriate to aviation;

4) It must test language use in a broader context than the use of ICAO phraseologies alone.

2.9 ICAO Rating scale

The ICAO rating scale delineates six levels of language proficiency ranging from the Pre-elementary (Level 1) to the Expert level (Level 6) across six areas of linguistic description: pronunciation, structure, vocabulary, fluency, comprehension and interactions. The detail of these rating scales and the criteria are shown in Table 2.1 (ICAO, 2004: A-8, A-9):

There are a few unique characteristics required by ICAO concerning its rating scale. First, the score given in any criterion including the overall score given to any candidate must be in a full score, i.e. not in a decimal or plus/minus e.g. 3.5 or 3+. Secondly, "the final score for each test-taker should not be the average or aggregate of the ratings in each of the six ICAO language proficiency skills but the lowest of these six ratings" (ICAO, 2008:19). It means that the overall score would be considered from the lowest score among all six criteria. Thirdly, its Level 4 is considered as "the safest minimum proficiency skill level determined necessary for aeronautical radiotelephony communications", hence, "a lower score than 4 for any one skill area indicates inadequate proficiency" (ICAO, 2008:19). Consequently, Level 4 is considered as „the cut-off score" since pilots who acquire any lower score than Level 4 would not be permitted by the Thai Department of Civil Aviation to conduct their flight operations on international flight routes (translated from Thai DCA announcement dated 16 February, 2010). These unique characteristics are likely to affect the way raters award scores to test-takers. The table below shows the details of the ICAO Language Proficiency Rating Scale.

Table 2.1: ICAO Language Proficiency Rating Scale

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	COMPREHENSION	INTERACTIONS
EXPERT 6	Pronunciations, stress, rhythm, and intonation, though possibly influenced by the first language or regional variation, almost never interfere with 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.	Comprehension is consistently accurate in nearly all contexts and includes comprehension of linguistic and cultural subtleties.	Interacts with ease in nearly all situations. Is sensitive to verbal and non-verbal cues, and responds to them appropriately.
EXTENDED 5	Pronunciation, stress, rhythm, and intonation, though influenced by the first language or regional variation, rarely interfere with 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 is accurate on common, concrete, and work related topics and mostly accurate when the speaker is confronted with a linguistic or situational complication or an unexpected turn of event. Is able to comprehend a range of speech varieties (dialect and/or accent) or registers.	Responses are immediate, appropriate, and informative. Manages the speaker/listener relationship effectively.
OPERATIONAL LEVEL 4	**Pronunciation, stress, rhythm, and intonation are influenced by the first language or regional variation, but only sometimes interfere with 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 is mostly accurate on common, concrete, and work related topics when the accent or variety used is sufficiently intelligible for an international community of users. When the speaker is confronted with a linguistic or situational complication or an unexpected turn of events, comprehension may be slower or require clarification strategies.**	**Responses are usually immediate, appropriate, and informative. Initiates and maintains exchanges even when dealing with an unexpected turn of events. Deals adequately with apparently misunderstandings by checking, confirming or clarifying.**
PRE-OPERATIONAL	Pronunciation, stress, rhythm, and intonation are influenced by the first	Basic grammatical Structures and sentence patterns associated with	Vocabulary range and accuracy are often sufficient to communicate on	Produces stretches of language, but paraphrasing and pausing are often	Comprehension is often accurate on common, concrete, and work related	Responses are sometimes immediate, appropriate, and informative. Can

3	language or regional variation, and frequently interfere with understanding.	predictable situations are not always well controlled. Errors frequently interfere with meaning.	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.	inappropriate. Hesitations or slowness in language processing may prevent effective communication. Fillers are sometimes distracting.	topics when the accent or variety used is sufficiently intelligible for an international community of users. May fail to understand a linguistic or situational complication or an unexpected turn of events.	initiate and maintain exchanges with reasonable ease on familiar topics and in predictable situations. Generally inadequate when dealing with an unexpected turn of events.
ELEMENTARY 2	Pronunciation, stress, rhythm, and intonation are heavily influenced by the first language or regional variation, and usually interfere with understanding.	Shows only limited control of a few simple memorized grammatical structures and sentence patterns.	Limited vocabulary range 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.	Comprehension is limited to isolated, memorized phrases when they are carefully and slowly articulated.	Response time is slow, and often inappropriate. Interaction is limited to simple routine exchanges.
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.

Note. – The Operational Level (Level 4) is the minimum required proficiency level for radiotelephony communication. Levels 1 through 3 describe Pre-elementary, Elementary, and Pre-operational levels of language proficiency respectively, all of which describe a level of proficiency below the ICAO language proficiency requirement. Level 5 and 6 describe Extended and Expert levels, at level of proficiency more advanced than the minimum required Standard. As a whole, the scale will serve as benchmarks for training and testing, in assisting candidates to attain the ICAO Operational Level (Level 4).

2.10 Raters and Factors affecting their rating

"Performance assessment necessarily involves subjective judgments" (McNamara, 1996: 117). These judgments involve acts of interpretation on the part of raters. In a subjective assessment like in writing and speaking tests, raters are one of the „facets" or main sources of variability in the scores (McNamara, 1996). There are three main sources of variability in the scores obtained when assessing a group of test takers. First, it is the relative ability of the test takers that differs unless the test involves a simple task within the competence of all test takers or a difficult one beyond every test taker"s competence. Second, it is the choice of task which the test takers choose. Finally, they are the raters who may give different scores for the same performance of the same test taker. This variability associated with raters is extensive and must be dealt with to derive stable and fair assessment. It means that the outcome of the test-takers" scores partly depend on raters. Many research studies concerning raters focus on the characteristics of raters in terms of their effects on the scores awarded to test-takers. One of these characteristics is rater"s bias.

Test bias is defined as "any aspect of a test which yields differential predictions for groups of persons distinguishable from each other by a factor which should be irrelevant to the test (Mousavi, 1999: 397). Candidates" age, their genders, their global/overall attitudes, and their nervousness are all irrelevant to the test. Raters must not consider these factors in their ratings otherwise, they will be biased. However, Wigglesworth (1993: 305) stated that the language assessment, particularly speaking and writing, is subjective and "it is subjected to the idiosyncratic differences which are found across raters". This idiosyncrasy is arduous to eliminate even after receiving rater training as McNamara (1996: 118) said, "rater differences are reduced by training but do persist".

McNamara (1996) explains that raters may differ from one another in many ways. First, raters may differ in their overall harshness/leniency. Secondly, raters may display particular patterns of harshness or leniency in relation to only one group of test takers. They may have tendency to overrate or underrate a test taker or a group of test takers. This is called „rater-test taker interaction". A rater may also be consistently lenient on one test item

while consistently severe on another. This is called „rater-item interaction". Thirdly, raters may differ from each other in the way they interpret the rating scale they are using. The problem arises because rating scales usually involve discrete rating categories, for example ICAO level 1 to level 6. When a test taker's ability falls roughly at the intersection of two levels, for example, above level 3 but still below level 4. The rater is forced into an „either/or" judgment at this point. One rater may decide to give level 3 to the test taker while another rater may decide to give level 4. Finally, raters may differ in terms of their consistency. This leads to another major concern in tests of speaking proficiency such as the ICAO English language proficiency test, which are subjectively scored. It is called „intra-rater" reliability which is "the degree to which an examiner or judge, making subjective ratings of ability, gives the same evaluation of that ability when he or she makes an evaluation on two or more different occasions" (Richards and Schmidt, 2002:273-4). This can have serious consequences for the candidates concerned, especially those in high stakes tests as pilots and air traffic controllers in English language proficiency testing. Another term concerning the consistency between raters is „inter-rater" reliability which is expressed as "the level of consensus between two or more independent raters in their judgments of candidates" performance." (Davies et al, 1999: 88).

In view of Emery (2006), raters are inevitably influenced by the many factors e.g. the rater"s first language, and if the raters are non-native speakers, the level of English language proficiency may have effects on them. Besides, if the raters have professional background in language and linguistics, it may depend on their degree of familiarity with aviation operations and aeronautical communication. On the other hand, if the raters have professional background in aviation operations, it may depend on their degree of familiarity with language and linguistics. Moreover, their degree of experience in language assessment and using language descriptors, the degree of training in the application of the rating scale, the extent and frequency of exposure to international accents and the extent and frequency of exposure to a particular accent could all have impacts on raters.

Moreover, results from previous studies (ibid) show that rater behavior and response vary with different groups in ways that can be partially attributed to variables such as

professional, cultural and linguistic background, extent of training in the use of assessment instruments, gender, amount of exposure to L2, and disparate and external pressures.

There are two experiential features that appear to be particularly salient i.e. effect of language experience and effect of professional experience (Shaw and Weir, 2007). Rater‟s language background is influential in terms of rater behavior and values. In writing assessment, raters who are acquainted with L1 rhetorical patterns show a tendency to be more sympathetic to L2 compositions, manifesting identical patterns unlike raters who are less familiar with these patterns (Hinkel 1994, Kobayashi and Rinnert 1996, Land and Whiteley 1989 cited in Shaw and Weir, 2007).

Professional experience between subject specialists and language-trained EFL teachers demonstrate a prominent effect in LSP testing. These two different kinds of raters tend to employ rating instrument differently (Elder, 1992). Brown (1995) developed an occupation-specific language performance test. She found that there were no overall differences between raters with linguistic background and raters with occupational experience in terms of grades awarded to candidates‟ performance. However, there were group differences in terms of the application of individual assessment criteria. This coincides with Hamp-Lyons‟ observation that EFL teachers attended to rhetorical criteria whereas the specialists emphasized content (Hamp-Lyons, 1991:134).

Leung and Teasdale‟s findings (1996) indicate that teachers-as-raters draw upon a range of professional experience, personal interpretations and folk theories in arriving at judgments in assessment. From teachers‟ perspective, they employ a range of issues as being importance in their assessment of their students such as the progress a student made, the age of a particular student relative to the rest of the peer group, the emotional state of the student, including home, cultural and linguistic factors that affect the student‟s performance. Such factors crucially affect provision as essential contextual information by which assessments can be interpreted. Some teachers in Leung and Teasdale‟s study even regarded the rating scales as having „little usefulness and little meaning‟ (Leung and Teasdale, 1996:66). They preferred relying on their own resources that built up over time from their own experience for

assessment. Shaw and Weir (2007:171) summarize “In terms of rater background, it appeared to be the case that different experiential backgrounds can affect the way in which markers assess, despite the fact that special training has been given to the rater for the specific marking exercise”.

Besides, rater expectations also show an effect on overall rater judgment (Weigle, 2002). Another issue is the interaction between the rater and the task difficulty. Weigle et al (2000) mention that raters may attempt to compensate for perceived task difficulty in applying the rating.

The group effects on rater reliability must also be considered. It has long been known in psychology that group dynamics can influence individual judgments (Shaw and Weir, 2007). Freedman (1981) argues that examiners could be trained to be more or less severe in their judgments. It has been found that examiner behavior varies with different groups, such as professional background, subject specialism and gender (Hamp-Lyons, 1990). This is due to each group having a unique frame of reference. This concurs with Brown‟s study (1995) which suggests that norms of judgment can be formed at the question level within tightly knit groups.

Rating conditions such as setting may additionally have effect on rater performance. Shaw and Weir (2007) suggest that familiarity with one‟s work conditions may result in a more settled and therefore less erratic performance. The „On site‟ marking where ratings take place at the test venue and „At home‟ marking where ratings are done at the raters‟ residence may affect the scores awarded by the same raters. Other variations in physical setting such as the provision of air conditioning (or heating) where the climate requires it or the presence of noise may also have an effect on the rating process. The temporal aspect as the time spent by raters may have an impact on the reliability of scoring (Vaughan, 1991). Raters who reach a decision quickly and stick to it tend to be more internally consistent raters than those who take a long time and vacillate (Shaw and Weir, 2007).

Another crucial factor which influences the manner in which the raters evaluate the test-takers‟ performance is the characteristics of the raters themselves (Shaw and Weir, 2007). Those characteristics are physical/physiological, psychological and experiential. The physical/physiological are short term ailments such as toothache, cold; long term ailment such as speaking, hearing, vision; age; and gender. In the case of gender, there is some evidence that gender plays a role. Not only have male and female raters been found to rate differently, but also test takers have been seen to respond differently to male/female interlocutors in the case that the interlocutor and the rater was the same person (Sunderland, 1995; Porter, 1991).

Another physical condition that may affect raters, particularly operational raters, is fatigue after flight duties. According to the Duty Regulations for Crew Members (Thai Airways, 2009) that normally requires a minimum of 24 hour rest period for crew members after their flight duties, any crew who gets rest period less than 24 hours is considered „not having enough rest‟ and it may make him fatigued. The same rule may apply to those operational raters who return from their last flight less than 24 hours and have to perform the duty as raters that it may affect their ratings due to their fatigue. This factor has not been studied concerning its effect on rating since this might be the first time that pilots are used as operational raters. Therefore, this issue needs further empirical investigation.

The psychological factors are rater personality, memory, cognitive style, affective schemata, concentration, motivation, and emotional state (Shaw and Weir, 2007). Raters‟ concentration and emotional state may be affected by the lack or inadequacy of sleep. The consequences for lack of sleep are far more dramatic than being tired in the morning. It can cause drowsiness or even headache. Sleep deprivation can result in impairment in cognitive function, such as attention, concentration and memory. Lack of sleep can cause mood swings including feeling low or being irritable (Ledoux, 2008). Not getting enough sleep can affect the ability to stay awake during the day or make raters feel fatigued. Life style can also have a huge impact on sleep and sleep quality. For example, irregular bedtimes and wake times might give rise to sleep problems that contribute to sleep deprivation. Operational raters who are pilots flying to different time zones could experience this irregular bedtimes and wake

times which may affect their duties as raters. Hence, it is worth considering this when assigning pilots to perform duties as operational raters.

Shaw & Weir (2007) mentioned that there has not been any empirical research study concerning the effect of factors associated with the environment or physical setting of the rating process on rater performance. This physical setting could be familiar or unfamiliar to raters e.g. if they are assigned to do their ratings at their office, at home, or some other preferred places. Shaw & Weir (ibid.) stated, "Familiarity with one‟s work conditions may result in a more settled and therefore less erratic performance." The provision of air conditioning or the presence of noise could also affect raters. This may have effect on the scores they award to candidates.

The experiential aspects are rater‟s education, rater‟s rating preparedness, rater‟s rating experience, rater‟s communication experience, and rater‟s first language, rater‟s familiarity with the target language, rater‟s target language competency, etc. Experience in rating also plays an important role in rater judgment. Shaw and Weir (2007:173) state that, in writing assessment, which is also a subjective assessment, stronger, experienced examiners appeared to attend less to the analytical activities and spend more time gaining an overall impression of the composition. The weaker, less experienced examiners attended more frequently to analytical activities. They also tended to be more positive in their comments.

Besides, some previous studies show that rater behavior and rater response varies with different groups in ways that can be partially attributed to variables such as professional, cultural and linguistic background, extent of training in the use of assessment instruments, gender, amount of exposure to the target language, and disparate and external pressure i.e. circumstantial, emotional, and psychological (Vann et al, 1991; Hamp-Lyons, 1990). Language background is particularly influential in terms of rater behavior and values. In writing assessment, raters conversant with first language rhetoric patterns undoubtedly demonstrate a tendency to be more sympathetic to L2 compositions, manifesting identical patterns unlike raters who are less familiar with these patterns (Kobayashi & Rinnert, 1996; Hinkel, 1994; Land & Whiteley, 1989). Effect of professional experience also plays an

important role in assessment. Comparisons are often made between how language proficiency exam raters and subject specialists rate. Subject specialists and language-trained teachers demonstrate a tendency to employ rating instruments differently (Elder, 1992). There are group differences between them in terms of the application of the individual assessment criteria (Brown, 1995). Brown (1995) who investigated rater background factors in assessment on the Japanese Language Test for Tour Guides, an advanced level occupation-specific oral test designed to measure the Japanese language skills of Australian Japanese-speaking tour guides and the intending tour guides, argues that „had the different groups been allowed to develop their own tests they might have been very different". It means that norms of judgment can be formed at the item level within homogeneous groups. Brown's results show that the occupational background of raters (with and without industry experience) add no bearing on the degree of consistency or the overall harshness of raters. However, teachers were harsher on three of language-related criteria i.e. grammar and expression, vocabulary, and fluency than industry raters, whereas industry raters were harsher on the criterion of pronunciation. This, somewhat, coincides with the study of Hamp-Lyons (1991) in her study of writing assessment which observed that EFL teachers „attended to rhetorical criteria foremost", whereas the specialists emphasized content (1991:134).

Elder's research finding (1992:15) states that "it is quite conceivable that in assessing use of subject specific language the ESL teachers are focusing on the lexis, grammar and the internal cohesion of the presentation while the subject specialists are more concerned about the way in which subject content is conceptualized." This finding offers the same evidence as Hadden (1990), Barnwell (1989), Ludwig (1982), and Galloway (1977) that language experts, whether they are teachers or trained language testers, have different perspectives of second language performance from other „linguistically naïve" native speakers.

However, Lumley (1995) found that ESL teachers and medical practitioners have broad similarities in judgments. They were somewhat lower for nurses and EFL teachers in medical interviews, where the major concern for the medical personnel was with their ability to give accurate information (Meldman, 1991).

Test task difficulty is another factor affecting the scores awarded to test-takers. In Generalizability theory (G-theory), task is considered as a factor or „facet" for specifying and estimating the relative effects of different factors on test scores (the other facets are raters and test-takers) (Upshur & Turner, 1999; Bachman et al., 1995; Bachman, 1990; Brennan, 1983; Cronbach et al., 1972). Raters" perspectives on test task difficulties may have effect on their decision-making since „easy" tasks may cause them to be harsher than usual. On the other hand, „difficult" tasks may make raters to be more lenient.

Rater leniency/severity (harshness) is a result of rater bias which is another factor affecting the rating. They are the attitudes shown by a rater towards a test taker"s performance (Davies et al, 1999). Some raters may be consistently generous (lenient), always giving relatively high scores to test takers; others may be consistently harsh (severe), giving relatively low scores; alternatively, raters may show bias towards or against particular groups of test takers. Differences in severity between individual raters will increase error associated with test scores and hence reduce their reliability (intra-rater reliability). Likewise, severity differences may occur between raters which affect the inter-rater reliability. McNamara (1996) states that raters may display particular patterns of harshness or leniency in relation to only one group of test takers, not others, or in relation to particular tasks, not others. That means there may be an interaction involving a rater and some aspects of the rating situation. Leniency or severity may not always work in the same direction for all items, or all things being rated. For example, raters in a speaking test may be asked to assess in three different criteria i.e. intelligibility, fluency, and accuracy. This rater may differ in the way he/she rates these criteria. He/she, who overall is fairly lenient, may be harsher when rating intelligibility. This is a kind of rater-item interaction which a rater is consistently lenient on one item while consistently severe on another. Another kind is rater-candidate interaction which a rater has a tendency to overrate or underrate a test taker or a particular group of test takers. Rater harshness is one of the factors concerning rater characteristics, which affect the scores awarded to test-takers. In Generalizability theory (G-theory), rater harshness is considered as a factor or „facet" for specifying and estimating the relative effects of different factors on test scores (the other facets are test tasks and test-takers) (Upshur & Turner, 1999; Bachman et al., 1995; Bachman, 1990; Brennan, 1983; Cronbach et al., 1972). McNamara (1996)

mentioned that raters may simply differ in their overall harshness/leniency, or they may be consistently lenient on one item while consistently severe on another (rater-item interaction), or they may have a tendency to over- or underrate a candidate or group of candidates (rater-candidate interaction). Even rater training cannot eliminate the extent of rater variability in terms of the overall severity (McNamara, ibid).

Another factor affecting raters in assessing speaking ability is the way they interpret the rating scale they are using. The problem arises because rating scales usually involve discrete rating categories (McNamara, 1996). When a test taker‟s ability of a speaking proficiency test falls roughly at the intersection of two of these rating categories, the rater is forced into an „either/or‟ judgment. One rater may consistently score such test taker with the higher level or category while another rater may consistently score the other way around.

A rating scale descriptor is “a statement which describes the level of performance required of candidates at each point on a proficiency scale” (Davies et al., 1999:43). In theory, raters refer to a rating scale in order to select a score to represent the candidate‟s ability in the trait of interest (Upshur & Turner, 1999). In reality, each rater has a unique background that may affect his/her judgment (Brown, 1995; Elder, 1993). Interpretation of a rating scale is always an interest of many researchers. Lumley (1995) found differences in the interpretation of the rating scale used by trained ESL raters and medical practitioners. This finding confirmed Brown‟s study about the perception of language-trained raters and experienced guides in 1995 that the two groups interpreted different criteria in different ways. Brown‟s conclusion of her study is interesting. She remarked that “raters appear to have inbuilt perceptions of what is acceptable to them and these perceptions are formed to some extent by their previous experience” and “it appears that even the explicitness of the descriptors and the standardization that takes place in a training session cannot remove these differences” (Brown, 1995: 13). The possible implication of this remark is that if the descriptors are inexplicit, raters‟ perceptions are prone to base on their previous experience. Imprecise rating scales often results in holistic marking by raters (Weigle, 2002 cited in Knoch, 2009). That leads raters to use the overall or global impression of the candidates in their ratings instead of using an analytic rating scale, as it should be (Knoch, 2009). This

misuse has significant effect in this ICAO proficiency assessment since ICAO requires the lowest score in any criteria to be the overall score (ICAO, 2004).

Being a native or non-native speaker of the target language is another issue that may affect raters‟ assessment. Barnwell (1989) suggests that native speakers are stricter raters. Van Maele (1994; cited in Elder *et al.*, 2001) found native raters attached far less importance to grammar than non-natives. Native raters were found to be more tolerant than non-native of grammatical inaccuracies and weak pronunciation when English was communicative. While register and intonation were largely peripheral to the non-native raters, they were central to native raters. In addition, non-native raters have been seen to adhere more closely to the established rating criteria while natives are more likely to be influenced by an intuitive feeling not captured by the descriptors (Brown, 1995).

It is likely that the interviewer‟s behavior during the interview will have some effect on the interview itself and consequently on the ratings themselves (Reed & Cohen, 2001). It has been suggested that unequal interlocutor support may well lead to bias in ratings (Lazaraton, 1996). Level and type of questions have, for instance, been found to influence ratings of the very same test taker when interviewed by different interviewers (Reed & Holleck, 1997). Likewise, over-accommodation to lowest-proficiency test takers in an interview situation may diminish the power of the probe and may also subsequently bias the ratings (Ross & Berwick, 1992). Many researchers have studied the roles of interlocutors in speaking assessment (Brown, 2003; Malvern & Richards, 2002; Jennings et al., 1999; McNamara & Lumley, 1997; Lazaraton, 1996; Ross & Berwick, 1992). Most of the foreign-language speaking assessment uses the oral proficiency interview technique which was developed by ACTFL (American Council on the Teaching of Foreign Languages) OPI proficiency scales out of the FSI (Foreign Service Institute) levels of oral proficiency (ACTFL,1999). This kind of interview technique was criticized concerning the „asymmetric nature‟ of interlocutor/candidate discourse (Taylor, 2000). It is the interlocutor who leads and controls the interaction during the interview. This creates the imbalance in the power relationship between interlocutor and test-taker. However, the effect of the interlocutor in this kind of assessment is undeniable. Various studies show how the behavior of the interlocutor

can affect candidate performance (Brown, 2005, 2003; O"Sullivan, 2000; Ross & Berwick, 1992). Brown (2004, 2003) and Brown & Hill (1998) found that raters" perception of a candidate"s oral proficiency, which affected the scores they awarded, was influenced by the choice of the interviewer.

A study based on data from the IELTS Oral Interview showed that interviewer styles and candidate styles can interact in ways that make it difficult for raters to distinguish the candidate"s talk from the interviewer"s talk (Brown, 1998). For example, an interviewer claiming personal knowledge of a topic, as opposed to mere interest, might take away a candidate"s reason for explaining. In this situation a rater would not be able to assume that a scant response by a candidate indicated lack of ability to elaborate.

The relationships between raters and test-takers can influence the way they award the scores. Bernardin & Buckley found out the „negative appraisal situation" for raters that they may be reluctant to „play god" hence leading to the tendency to be lenient as defensive behavior i.e. avoiding the reactions from candidates who are someone with personal relationship by not awarding harsh rating (Bernardin & Buckley, 1981: 209). Moreover, Papageorgiou"s finding, which was conducted with a group of 12 expert judges, revealed that their decision-making was affected by expectations due to „bias of insiders" because the judges knew the test-takers (Papageorgiou: 2010). It means that the judges did not refer to the scales used in their assessment but considered examinees they knew or examined. This kind of personal relationship between raters and test-takers affects raters" decision-making.

Finally, there is the issue of the length of rater training and its nature. Davies et al (1999:161) explain that rater training is the preparation of raters for their task of judging performances. During this kind of training, it often takes the form of workshop in which raters are acquainted with the test format, test tasks, and the rating criteria. Exemplar performances at each defined level of performance are presented and discussed. In rater training, raters are introduced to the assessment criteria and asked to rate a series of carefully selected performances, usually illustrating a range of abilities and characteristics issues arising in the assessment. Ratings are carried out independently. Raters are asked to evaluate

a series of performances, to compare their ratings, and to discuss any differences between them. Raters are shown the extent to which they are in line with other raters and thus achieving a common interpretation of the rating criteria. Subsequently, raters may be asked to rate a further set of performances. The rating session is usually followed by additional follow-up ratings in order to determine if the rater can participate satisfactorily in the rating process (McNamara, 1996:125-126). Only those raters reaching a predetermined level of conformity with the generally agreed ratings are certified as raters.

The fairness of an assessment involves the use of raters who have been trained carefully in the use of the rating procedure, and who have demonstrated a required level of agreement with the raters in moderation sessions and the practice of rating each script more than once, and the adoption of procedures for dealing with disagreement, such as averaging ratings, getting a further rating, or bringing the raters together to reach agreement.

Studies of the effect of rater training show that training reduces extreme differences in severity between raters and makes raters more internally self-consistent, but significant differences in severity between raters remain (Davies *et al.,* 1999). Rater characteristics such as relative severity and self-consistency vary over time. Trained and untrained raters have been shown to disagree on scale points (Barnwell, 1989). Halleck‟s study (1996) of certified OPI raters and trainees found that they agreed more on some levels (superior and intermediate mid) than on others (advanced high, advanced, and intermediate high). Thus, rater training is one of the factors affecting raters in assessment. As cited by McNamara (1996:126), McIntyre (1993), Weigle (1994) and Shohamy et al (1992) conduct the research about the effectiveness of rater training and demonstrate that rater training is successful in making raters more self-consistent. Weigle (1994) also states that reliable measures are unlikely to be achieved from untrained raters. Rater training can reduce, but by no means, eliminate the extent of rater variability in terms of overall severity. Lunz & Stahl (1990 cited in McNamara, 1996:126) argue that raters employ unique perceptions which are not easily altered by training. However, it is usually required to have two or more raters who are trained to agree on independent ratings of the same performance. However, Weigle (1998) stated that a focus on rater consensus may compel raters to ignore their own expertise and

experience in assessing. Therefore, it is important that rater training must not force raters to overlook their own expertise and experience in their decision-making. As McNamara contends that the traditional objective of rater training which was to eradicate any differences between raters may be „unachievable and possibly undesirable" (McNamara, 1996:232), on the other hand, he argues that the more desirable aim of rater training is to get raters to become more focused and to encourage raters to be self-consistent.

In summary, rater training can bring raters" differences in severity to a tolerably acceptable level but it cannot totally eradicate differences in severity. It can make raters more consistent in their individual approach to scoring. It may be said that leniency and severity are fixed traits of raters. Raters display certain characteristics in their participation in the rating process. These characteristics are a source of potentially considerable variability in rating performances. Rater training is essential for creating the conditions for an orderly measurement process based on ratings by making raters more self-consistent. The most appropriate aim of rater training is to make raters internally consistent.

2.11 Content analysis

The term „ontent analysis" is defined by Mousavi (1999: 61) as "a general term covering a variety of methods for analyzing a discourse, message or document for varying themes, ideas, emotions, opinions, etc. Most such analyses consist of sophisticated counting schemes in which the frequency of particular words, phrases, affective expressions and the like are determined." Richards & Schmidt"s definition is quite similar to Mousavi"s as "a method used for analyzing and tabulating the frequency of occurrence of topics, ideas, opinions and other aspects of the content of written and spoken communication" (Richards & Schmidt, 2002: 114).

However, George (1959) introduced a different distinction of content analysis, which focuses on the aspects of the communication content from which the analyst draws inferences regarding non-content variables. George (1959) classified content analysis as two approaches, quantitative and non-quantitative content analysis. Quantitative content analysis

is concerned with the frequency of occurrence of given content characteristics while non-quantitative content analysis makes inferences from content to non-content variables. This approach needs not be based on the frequency values of content features. It uses „non-frequency" content indicators such as presence or absence of a given content characteristic for the purpose of inference. This non-quantitative or non-frequency approach utilizes the mere occurrence or non-occurrence of attributes for purposes of inference (George, 1959: 145). George emphasized that the non-frequency approach was a "more conventional way of interpreting communication and drawing inferences" from the content. This non-frequency approach is particularly difficult to objectify since it requires considering the situational, behavioral, and linguistic contexts into account.

Hsieh & Shannon (2005: 1278) have also defined „non-quantitative" content analysis or „qualitative" content analysis as "a research method for the subjective interpretation of the content of text data through the systematic classification process of coding and identifying themes or patterns". It can be seen from these definitions that qualitative content analysis goes beyond simply counting words or extracting objective content from texts to examine meanings. It is designed to explore the meanings underlying physical messages. As Graneheim & Lundman (2004: 106) say "reality can be interpreted in various ways" and "the understanding is dependent on subjective interpretation", qualitative content analysis plays an important role in this kind of interpretation. Zhang & Wildemuth (2009) emphasize that qualitative content analysis pays attention to unique themes that illustrate the range of the meanings of the phenomenon rather than the statistical significance of the occurrence of particular texts or concepts.

A basic issue when performing qualitative content analysis is to decide whether the analysis should focus on manifest or latent content (Graneheim & Lundman, 2004). The manifest content is the visible and obvious components of the text while the latent content involves an interpretation of the underlying meaning of the text (Downe-Wamboldt, 1992; Kondracki et al., 2002). Both manifest and latent contents deal with interpretations but the interpretations vary in depth and level of abstraction.

The steps in qualitative content analysis are to select a „unit of analysis“ and a „meaning unit“. The most suitable unit of analysis, which was suggested by Graneheim & Lundman (2004), is whole interviews while a meaning unit can be words, sentences or paragraphs containing aspects related to each other through their content and context. The label of a meaning unit is referred to as a „code“. A code can be assigned to discrete objects, events and other phenomena. Rubin & Rubin (1995: 238) define coding as "the process of grouping interviewees“ responses into categories that bring together the similar ideas, concepts, or themes". Coding can be in the forms of names, evidence, time sequences, hesitations, signs of emotion, indications of fear or amusement, etc. It can be analyzed as length of pauses, the order of wording, and the exact words that were used. According to Rubin & Rubin (ibid), anything can be coded to help analyze the data.

Watzlawick et al. (1967: 66) stated, "Human beings communicate both digitally and analogically" which are explained by Graneheim & Lundman (2004:111) as, "Verbal communication is mainly digital and easily transcribed into a text while non-verbal communication is mainly analogical and often put at a disadvantage in the transcription process". Therefore, it is valuable to notice silence, sighs, laughter, postures, gestures, etc. as they may influence the underlying meanings (Graneheim & Lundman, ibid). This kind of non-quantitative or qualitative content analysis was used as a method to answer the fourth research question of this study.

The literature review in this chapter provides the contents and constructs of the questionnaire and the semi-structured interview. In addition, the responses obtained from the questionnaire and the interview transcripts of this study were analyzed based on the technique of content analysis presented in the previous section. The factors affecting the raters“ decision-making to be determined are listed below:

- Raters“ educational and rating backgrounds,
- Raters“ mental conditions,
- Raters“ physical conditions,
- Physical settings,

- Raters‟ rating strategies,
- Test tasks and speech samples,
- Interviewer/interlocutor effects,
- Candidates/test-takers,
- Rating scales and descriptors,
- Cut-off score,
- Personal relationship between raters and candidates,
- Scoring techniques, and
- Raters‟ harshness/leniency.

This chapter reviews the English for Occupational Purposes in general, aviation English in particular, oral proficiency testing, and various factors affecting raters‟ decision-making. The above thirteen factors will be investigated and discussed in the subsequent chapters.

CHAPTER III

RESEARCH METHODOLOGY

This chapter presents the research methods and procedures used in this study. Five major areas covered in this chapter are research procedures, subjects, research instrumentation, data collection and data analysis.

3.1 Research procedures

The mixed method approach involving combinations of quantitative and qualitative research methods (Dornyei, 2007) was applied in this study. The quantitative approach was employed to investigate the relationships among the different backgrounds of raters, the relationships between trained and untrained raters, and the interactive relationships between rater backgrounds and their training on rating Thai pilots‟ speaking ability. While the qualitative approach was employed to examine the factors affecting the decision-making of the raters with different backgrounds and training in rating Thai pilots‟ English speaking proficiency.

3.2 Participants

The participants in this study were classified into two groups: the test taker group and the rater group.

3.2.1 The test taker group

This study applied the purposive sampling technique to select the speech sample data. In the study, there were 10 pilots working for Thai Airways International PLC who took the RELTA. Since the participation of RELTA testing with Thai Airways International PLC was on a voluntary basis, there were only 11 pilots who volunteered to take part. One of them withdrew before the testing actually started. This is the reason why there were only 10

participants. These participants were divided into three groups, namely Level 3, Level 4, and Level 5 groups. These pilots were experienced line pilots, which mean that they had been working as airline pilots for at least one year, not pilots who just graduated from their flying school and did not have airline experience or “ab initio pilots”.

3.2.1.1 Source of data for the pilot study: One speech sample was randomly selected from Level 4 group by simply drawing lots from the pool of Level 4 speech sample. Then this speech sample was given to the group of four raters for the pilot study. These four raters were not included in the main study.

3.2.1.2 Source of data for the main study: The researcher randomly selected the other three speech samples – each from Level 3, Level 4, and Level 5 – by simply drawing lots from the pool of Level 4 and Level 5 speech samples. Since there was only one Level 3 speech sample, it was selected as the only availability in this study. These three speech samples were distributed to the raters considering the appropriateness of the time in rating that the participants were convenient. These three different performances were selected as the speech samples so that the mixed levels of proficiency would be presented in the assessment. These speech samples were distributed to 20 raters for the main study.

3.2.2 The rater group

This group included linguistic raters and operational raters. The purposive sampling technique was applied to this group based on their willingness to participate and their availability at the time of the study.

3.2.2.1 Participants for the pilot study: The participants in this group comprised four raters, one linguistic rater with rater training experience, one linguistic rater without rater training experience, one operational rater with rater training experience, and one operational rater without rater training experience.

3.2.2.2 Participants for the main study: The participants consisted of 10 linguistic raters, five of them with rater training experience and the other five without, and 10 operational raters, five of them with rater training experience and the other five without. These raters did not participate in the pilot study.

The raters, both in the pilot and in the main study, were not informed of the levels of the speech samples they were given. They received only the instructions and the printed details about ICAO rating scales and descriptors. They did not obtain any kind of briefing or in-depth information concerning the ICAO requirements from the researcher. Then, they listened to the given speech samples, and rated them by using the given ICAO rating scales. Right after finishing their ratings, they answered questionnaires concerning their personal data such as their age, gender, educational background, rater training background, etc. After completing the questionnaires, they were interviewed by the researcher. Even though their age and gender were not the focus of this study, they were included in the questionnaire for the interest of any further study.

3.3 Data source

The following source of data was used in this study.

RELTA (RMIT English Language Test for Aviation)

RELTA stands for „The RMIT English Language Test for Aviation". RELTA is a standardized test developed by RMIT English Worldwide, a global English language learning institution based in Melbourne, Australia, which is a part of RMIT Training Pty Ltd, a wholly owned commercial subsidiary of RMIT University, which is one of Australia"s largest universities. RMIT is a world leader in aviation research and training, with dedicated schools in Aerospace Engineering, Flight Training and English language training and assessment. RELTA has been designed to allow the language proficiency of pilots and air traffic controllers to be assessed according to the ICAO Language Proficiency Requirements.

There are two streams of RELTA; one for pilots and the other for air traffic controllers whose first language is not English. The test has been designed to assess pilots or air traffic controllers against the six levels of ICAO Language Proficiency Scale. Both forms of RELTA have been specifically designed for existing pilots and air traffic controllers whose first language is not English, and whose language proficiency needs to be assessed for licensing purposes in line with the ICAO Language Proficiency Requirements taking effect in 2008. RELTA, which was used as the data source in this study, is an early version of the one for pilots that was conducted with a group of THAI pilots.

RELTA Pilot test, which is confidential, hence, not provided in this study, comprises two parts, a listening part and a speaking part, each part has three sections. Only the speaking part was employed in this study.

The speaking part starts with the "warm up" section, which takes approximately one minute. Its format is non-face to face. There are two questions related to candidates" background and four questions related to test contexts (visual information provided). The mode of delivery is computer-mediated with an interlocutor asking questions. This "warm up" section is not assessed, hence there is no mark given.

The speaking part section 1 follows the "warm up" section. It takes approximately five minutes. The candidates are required to produce language of which ICAO standard phraseology alone can convey the message. Its format is a direct/live non-face to face role-play in a continuous dialogue. Each candidate assumes the role of a pilot and interacts with a live interlocutor who assumes the role of an air traffic controller in the linear and continuous dialogue. The mode of delivery is computer-mediated with the interlocutor controlling the audible/visual. The interlocutor follows a prescribed role-play script contained in the Examiner Booklet. The language elicited for assessment is ICAO standard phraseology in simple familiar, routine and predictable situations in radiotelephony communication contexts. This section score is weighted 20%.

The speaking section 2 takes about six to eight minutes. The candidates are required to produce language of which both phraseology and plain English are required to convey messages (only responses where prompts have been designed to elicit plain English are assessed). Its format is also direct/live non-face-to-face role-play for continuous exchanges mirroring real-time communication. Each candidate assumes the role of a pilot and interacts with the interlocutor who assumes the role of an air traffic controller in the linear and continuous dialogue. The mode of delivery is computer-mediated with an interlocutor controlling audible/visual. The interlocutor follows a prescribed role-play script contained in the Examiner Booklet. Language elicited for assessment is Plain English in both complex non-routine and unpredictable radiotelephony communication contexts. This section score is weighted 35%.

The speaking section 3 requires around 10 to 12 minutes. The candidates are required to communicate in general English and relate to the concepts in Section 2 before expressing preferences and discussing abstract topics of which they offer opinions and speculate about the future. Its format is face-to-face interview. The mode of delivery is that the interlocutor asks prescribed questions, which are contained in the Examiner booklet. The language elicited for assessment in this section is general English in aviation specific contexts, of which the themes from section 2 are provided for discussion. This section is weighted 45%.

The last section is the closing/wrap-up. It briefly carries on for around 30 seconds. The format is still face-to-face but there is no language elicited for assessment. In the same way as the “warm up” section, there is no score given for this concluding section.

Scoring

Regarding the scoring of RELTA, it requires the application of the ICAO six-band language proficiency rating scale (see Table 2.1, page 37). In the process of test administration, live examiners and human raters are employed to ensure all aspects of the ICAO requirements are applied in the delivery and rating of RELTA. All rating and reporting processes are externally controlled to ensure security, fairness and accountability (this includes double and triple rating of speaking performances to guarantee fairness and

accuracy of results). Comprehensive examiners and rating programs are provided to ensure test delivery and rating is accurate for all candidates. At least two trained and qualified RELTA raters are used to determine a candidate‟s scores and ICAO level to maximize fairness and accuracy of results. RELTA claims to be practical and easy to administer, with simple and efficient pre-test, in-test and post-test administration procedures in place to facilitate security, test delivery and reporting efficiency.

RELTA reliability & validity

RMIT English Worldwide also conducted RELTA validation to confirm its reliability and validity. All forms and versions of RELTA have been extensively trialed and validated with actual target-user candidate populations. Trials have been conducted with over 150 non-native speaker aviation personnel. Therefore, they have been found to be valid and reliable through extensive research and statistical analysis. Test trials conducted to date on all versions of RELTA for Pilots indicate that both the listening and speaking components of RELTA produce consistent scores. The listening test has reliability coefficients ranging from .76 to .90. Reliability is established by determining Cronbach‟s alpha. Inter-version consistency is verified through concurrent validation to establish score-equivalence between versions. This is determined by computing correlation coefficients (Pearson‟s Product Moment Coefficient). All forms of the speaking component of RELTA for Air Traffic Controllers were found to be reliable providing effective rater training and application of rating processes occurs. The listening component was also found to be reliable, with a reliability coefficient of .70 or above. As the reliability of the Speaking test scores is contingent on high intra- and inter-rater reliability, RMIT English Worldwide (REW) provides quality checks to monitor rater reliability and provide ongoing recurrent rater training.

Concerning its construct validity, while RELTA is an ESP test in radiotelephony communication and requires extensive background knowledge, data analysis indicates a good positive correlation between the Listening and Speaking test and TOEIC Test (extremely high correlations are not necessarily expected, since the TOEIC and RELTA tests are designed to measure different language skills and different language domains).

Extensive research and presentation of data in the form of a Master"s degree thesis in Language Testing: "The Development of an ESP Proficiency Test for Civil Airline Pilots: Investigating Construct Validity" supports the overall RELTA test construct (Kay, 2005). The research indicates that RELTA is a valid test for aviation personnel, and is extremely effective in assessing proficiency in both phraseology and plain English in a range of work-related communicative contexts. This research also confirms that RELTA is an effective proficiency test in assessing candidates at all six ICAO proficiency levels, and for all six ICAO criteria. Furthermore, it is established that RELTA is an effective proficiency test in the language domains relevant for pilots/controllers, supported by the fact that general English proficiency tests are not able to detect proficiency levels according to the ICAO standards when administered to the same test trial populations. Findings also indicate that the assessment of proficiency in both phraseology and plain English is valid and appropriate. Pilots and controllers may be experienced and therefore competent in using phraseology, but lack proficiency in plain English. Similarly, there is a trend with less experienced personnel to occasionally be more proficient in face-to-face communication contexts, but lack communicative competence in radiotelephony communications.

In terms of content validity, RELTA has very high content validity, with Section 1 of the listening and speaking components behaving effectively for the assessment of communicative ability in routine phraseology; Section 2 effectively assessing plain English in non-routine radiotelephony and Section 3 assessing plain English in conversational contexts. In addition, extensive data analysis and application of different weighted scores for each of the three Speaking sections indicated that the 20/35/45% section weighting is effective in separating and evaluating candidates effectively over all six ICAO proficiency levels.

About the face validity and authenticity, feedback in the form of test trial evaluation surveys and focus groups among test-taker target groups indicates that RELTA has very high face validity. Following test trials, participants comment on the content of the test as being appropriate and related to their jobs. In addition, the images associated with the tasks (both

prompts and context-setting photographs) are perceived as providing a high level of authenticity, allowing candidates to interact and engage with the test effectively. For example, the most recent test trial in Korea reported 92%, 87% and 76%, respectively, of the trial participants found Section 1, 2 and 3 of the RELTA speaking to be relevant. In addition, 80% stated that the test trial was an appropriate assessment tool for their profession.

Focusing on the test tasks, each RELTA test section purpose is related to the six criteria of the ICAO Language Proficiency Requirements that reflects both the plain English language and operational knowledge of the pilots. Firstly, the use of phraseology in radiotelephony (voice-only) in RELTA speaking section 1 reflects the pilot‟s pronunciation in routine radiotelephony contexts, his/her use of range of vocabulary in phraseology, the ability to construct transmissions using phraseology, the fluency of phraseology in transmissions, the immediacy and appropriateness of responses and ability to check, confirm and clarify information and deal with misunderstandings using phraseology, and the comprehension of pilot/controller exchanges associated with phraseology in routine contexts. The knowledge of aircraft operating procedures and associated phraseology in routine situations is required, though it is not assessed.

Secondly, the use of plain English in radiotelephony (voice-only) in RELTA section 2 casts back the pilot‟s pronunciation of plain English in non-routine radiotelephony contexts, his/her use of range of aviation specific vocabulary, the use of grammatical range and accuracy of plain English in non-routine radiotelephony situations, the fluency of plain English in non-routine radiotelephony contexts, the immediacy and appropriateness of responses and ability to check, confirm and clarify information and deal with misunderstandings during non-routine radiotelephony events, and the comprehension of pilot/controller exchanges associated with non-routine events involving plain English. The knowledge of aircraft operating procedures and associated phraseology in non-routine situations is required as well, still it is not assessed.

Lastly, the use of plain English in conversation (face-to-face) in RELTA speaking section 3 throws back the pilot‟s pronunciation of plain English in face-to-face aviation

related contexts, his/her ability to produce language fluently and knowledge of discourse markers, the use of grammatical range and accuracy in work-related conversational contexts, the fluency of plain English in work-related conversational contexts, the immediacy and appropriateness of responses and ability to check, confirm and clarify information and deal with misunderstandings in conversational contexts, and the comprehension of plain English in a work-related conversational context. The knowledge of flight processes and issues in aviation is also required but not assessed.

3.4 Research instrumentation

The following instruments were used in this study.

3.4.1 Questionnaires for Raters

A questionnaire for raters (see Appendix C) was developed primarily from an extensive research of relevant literature and was designed to elicit the raters‟ personal information and opinions, then used with the participants. The questionnaire was divided into three main parts.

Part 1: There were nine items in this part. The participants were asked to choose and answer about their personal information such as their genders, age and educational background. Their experiences in rating were also included in this part. Even though their age and gender were not the focus of this study, they were included in the questionnaire for the interest of any further study.

Part 2: There were 52 items in this part. Part 2 had two sections. A Likert-type of questionnaire was developed to assess the raters‟ familiarity with various English accents, their familiarity with linguistics and aviation operations, their familiarity with the ICAO rating scale and descriptors, and their rating strategies. Participants were asked to respond on the five-point Likert scale ranging from „never‟ to „always‟ in this section. The next section in this part asked the participants concerning the factors that might affect their ratings in the

form of yes/no questions. The similar questions were repeated again to acquire the in-depth information during the interview.

Part 3: There was one open-ended question asking for anything the participants wanted to comment concerning the assessment.

The questionnaire validation process

Regarding the validation of the questionnaire, which was carried out in the pilot study, two main stages were performed:

1. Priori validation:

With regard to the preliminary study conducted, a Likert-type and yes/no questions were developed, and the first draft was used in the pilot study to elicit the raters‟ personal information and opinions. The whole questionnaire consisted of 62 items. For item number 1 to 9, the participants were asked to select the choices that fit them most and answer the questions concerning their educational levels, their first language, and the rater training program(s) they attended. Then, they were requested to respond to the five-point Likert scale ranging from never to always for item number 10 to 26. After that, they were asked to answer the yes/no questions for item number 27 to 61. The last item inquired any comment the participants wanted to make. All items in the questionnaire were written in English since all participants were expected not to have any problem understanding them. The questionnaire in the pilot version consists of three parts:

Part 1: Personal information (Choosing the answers and filling in the blank)	9 questions
Part 2 – Section 1: Familiarity with various English accents, familiarity with linguistics and aviation operations, familiarity with the ICAO rating scale and descriptors, and rating strategies	17 questions (5 choices)
– Section 2: Factors that might affect their ratings	35 questions (yes/no)

Part 3: Comments	1 question (open-ended)
Total	62 questions

The item-Objective Congruence (IOC) index was used to improve the content and construct validity. Three independent experts in the field of language and aviation, who matched each item with the specific domain to be observed, considered the IOC index. The criteria for selecting the experts were that all experts were related to the field of language and/or aviation as previously mentioned. They were two university level lecturers and an airline pilot working for an international airline.

2. Posteriori validation:

According to the judgment and comments of the three experts on the contents and constructs of the questionnaire, the results of the Item-Objective Congruence (IOC) index were presented as follows:

Regarding the content validity of the questionnaire, the experts highly agreed that the content of the questionnaire reflected the objectives of the questionnaire. However, one of the experts suggested that she wondered if it was necessary to define the terms „rarely", „sometimes", „frequently" and „always" because different people might interpret/perceive these terms differently. Furthermore, the same expert doubted whether the participants would answer the question of „rater harshness" truthfully. However, all three experts believed that the questionnaire was appropriate to investigate the personal information and opinions of the participants. They also strongly agreed that specific language used in the questionnaire could be found in real conversation when speakers encountered language difficulty. The format of the questionnaire was accepted by the experts that it was appropriate, straightforward, not too laborious, and not over-complex. Finally, there were some experts" comments e.g., the researcher should clarify the criteria of the scale „never" to „always" in terms of percentage. For example, „always" could be defined as the context exactly true for me (100%). The questionnaire seemed to explore what it claimed to explore, and thus it might be concluded that the content validity of the questionnaire was satisfactory.

With regard to the construct validation, the result of the IOC analysis indicated that the average IOC index of the questionnaire was 0.94. Additionally, 94.19% of the questionnaire items had an IOC index equal to or more than 0.67, which means that this 94.19% of the questionnaire items were accepted and were retained. That was because they were congruent with the objectives and could acquire what the questionnaire was intended to acquire. On the other hand, the rest of questionnaire items with the IOC index of less than 0.7 should be revised or rejected as they had unsatisfactory ability to assess what the questionnaire intended to measure. To sum up, the constructs of the questionnaire seemed, in general, to be accepted by the experts. Most of the items were retained and used in the main study and a few items were revised.

It could be concluded that both construct and content validity of the questionnaire were satisfactory, meaning that it could assess what it was intended to measure.

After the pilot study of the questionnaire had been done, the revised version of the questionnaire was developed.

3.4.2 Rater score sheet and remarks

A rater score sheet and remarks was provided to each rater in order to specify the scores given to each test taker in each criterion and the overall score, and to state the reasons why the rater awarded such scores to the test takers or other comments the rater would like to make.

3.4.3 Interviews

A series of semi-structured interviews was constructed from relevant literature similar to the questionnaire. Thereafter, face-to-face interviews were conducted to elicit their opinions concerning the in-depth information of the raters‟ strategies towards their ratings of the test takers‟ proficiency and other factors affecting their decision-making. Semi-structured interviews were added as they allowed a greater depth of meaning to emerge than by using questionnaires alone (Polit & Hungler, 1999). Because the interviews were semi-structured,

the exact interview questions varied from one rater to another. However, the interviews were controlled since all raters were interviewed by the researcher.

The interview validation process

The validation of the interview questions was also performed in the pilot study. The two main stages as similarly conducted with the questionnaire validation process were conducted:

1. Priori validation:

Most of the interview questions were similar to the questionnaire items and they were repeated during the interviews. The first draft was used in the pilot study to obtain the in-depth information concerning the factors affecting the raters‟ decision-making. The interview consisted of 58 questions. These questions were classified into 13 groups according to the 13 factors affecting raters‟ decision-making as described in Chapter 2. For question number 1 and 2, the raters were asked to describe their educational background at the university degree level and the relationship between their educational background and rating. The raters were asked about their mental conditions in question number 3 to 6. After that, they were called on to describe their physical conditions for question number 7 to 9. For question number 10 to 13, the raters were asked about the physical setting where they rated. Question number 14 to 27 touched on the raters‟ rating strategies. The raters‟ opinion concerning the test task and the speech samples were asked by the question number 28 to 32. Question number 33 to 35 inquired how the raters thought of the interview/interlocutor performance. How the raters felt about the candidates being elicited in question number 36 to 41. Question number 42 to 49 were asked to obtain the raters‟ perspectives toward the ICAO rating scale and descriptors. The raters‟ standpoint toward the cut-off score and its consequences including the candidates‟ pass/fail results were explored in question number 50 and 51 while question number 52 looked into the personal relationship between the raters and the candidates. Question number 53 and 54 scrutinized the ICAO scoring requirement and its effect toward the raters‟ score awarding. Question number 55 requested the raters to self-consider their harshness/leniency. The last three questions (number 56 to 58) sought other raters‟ opinions

i.e. the raters‟ utmost concern in awarding the scores, their ideal characteristics of a rater and any comments they might have about this research. All questions were prepared in English.

In conclusion, the interview question in the pilot version consisted of 58 questions as follows:

Question 1 and 2: Educational background	2 questions
Question 3 to 6: Mental conditions	4 questions
Question 7 to 9: Physical conditions	3 questions
Question 10 to 13: Physical settings	4 questions
Question 14 to 27: Rating strategies	14 questions
Question 28 to 32: Test tasks & speech samples	5 questions
Question 33 to 35: Interview/interlocutor	3 questions
Question 36 to 41: Candidates	6 questions
Question 42 to 49: Rating scale & descriptors	8 questions
Question 50 and 51: Cut-off score	2 questions
Question 52: Personal relationship	1 question
Question 53 and 54: Scoring	2 questions
Question 55: Rater harshness/leniency	1 question
Question 56 to 58: Others	3 questions
Total	58 questions

The item-Objective Congruence (IOC) index was used to improve the content and construct validity. The same three independent experts employed in the questionnaire validation considered the IOC index of the interview questions.

2. Posteriori validation:

Based on the judgment and comments of the three experts on the content and construction of the interview questions, the results of the Item-Objective Congruence (IOC) index were presented as follows:

Concerning the content validity of the interview questions, the experts highly agreed that the content of the interview questions reflected the objectives of the interview questions. However, one expert did not agree with the question concerning the physical condition regarding the hours of sleeping. He commented, "Each person"s sleeping habit is different and the number of hours may not indicate physical fatigue". Furthermore, another expert doubted whether the participants would answer the question of „rater harshness" truthfully. However, all three experts believed that the interview questions were appropriate to investigate the factors affecting the raters" decision-making. They also strongly agreed that specific language used in the interview questions could be found in real conversation when speakers encountered language difficulty. The format of the interview questions was accepted by the experts that it was appropriate, straightforward, not too laborious, and not complicated. The interview questions seemed to elicit what it was purported to elicit, and thus it might be concluded that the content validity of the interview questions was satisfactory.

With regard to the construct validation, the result of the IOC analysis indicated that the average IOC index of the interview questions was 0.96. Additionally, 96.55% of the interview questions had an IOC index equal to or more than 0.67, which means that this 96.55% of the interview questions were accepted and were retained. That was because they were congruent with the objectives and could assess what the interview questions were intended to obtain. On the other hand, the rest of interview questions with the IOC index of less than 0.67 were revised or rejected as they had unsatisfactory ability to elicit what the interview questions purported to elicit. To sum up, the constructs of the interview questions seemed, in general, to be accepted by the experts. Most of the interview questions were retained and used in the main study and a few interview questions were revised.

It could be concluded that both construct and content validity of the interview questions were satisfactory, meaning that it could elicit what it was aimed to assess.

After the pilot study of the interview questions had been done, the revised version of the interview questions was developed.

3.5 Data collection

The 10 test takers already took RELTA. Their speech samples were recorded and were given to the raters who listened to three speech samples and rated them according to the ICAO rating scale and descriptors given to them. The raters were also asked to state their given scores and the reasons and /or comments why they awarded such scores to the test takers in the provided sheets. Each rater was asked to complete his/her rating in a provided area.

Then the raters were asked to finish the rater questionnaires. After the rating was completed, each rater was invited to participate in a semi-structured interview. The interviews were conducted in a quiet room and lasted for 45-60 minutes. The interviews were administered in the language requested by each individual rater i.e. in Thai or English subjected to each rater‟s preference.

3.6 Data analysis

Four types of data analysis were undertaken. The test results, the rater questionnaires and the interviews were analyzed as follows:

3.6.1 Hypothesis testing

To answer the three hypotheses, the 2x2 ANOVA was employed.

3.6.2 The rater questionnaires

The content validity of the questionnaire was validated by three content experts. The construct validity was checked using the IOC index as aforementioned in 3.4.1.

The data from the rater questionnaires were presented in the form of tables as shown in Chapter 4 (see Table 4.15, 4.16, 4.17, 4.18).

3.6.3 The rater score sheets and remarks

The data from the rater score sheets and remarks were presented in the form of descriptive data as shown in Appendix F, G and H.

3.6.4 The interviews

The content validity of the interview was also validated by three content experts. The construct validity was checked using the IOC index as mentioned in 3.4.3 above.

The interview data were transcribed and translated into English - if the raters answered in Thai, then qualitative content analysis was conducted. The content from the interview was grouped into types reported in the literature and was analyzed by the qualitative content analysis technique.

The process of the qualitative content analysis of the interview data are as follows:

1. Prepare the data.

The recorded voice of the raters was transformed into the written text by transcribing it in order to reveal the information related to their behaviors and their thoughts. Only the main questions from the interview corresponding with the questionnaire were transcribed. The verbalizations were transcribed literally and the observation during the interview e.g. pauses, sighs, etc. were also noted.

2. Define the unit of analysis, theme & sub-theme.

The unit of analysis was the interview transcripts. The themes consisted of the 13 factors presented at the end of Chapter Two. If any theme comprised more than one topic, it was divided into sub-themes e.g. the theme of Educational & Rating background was divided into two sub-themes of Educational and Rating background.

3. Develop categories and a coding scheme.

The categories were divided into non-verbal and verbal expressions. The non-verbal expressions were signs of emotions such as hesitations, discomfort, uneasiness that were exhibited in the forms of pauses, sighs, silence. The verbal expressions were the manifest contents, which were visible and obvious components such as words expressed by the raters. These signs of emotions and verbal expressions were considered as meaning units referred to as content units or coding units.

4. Code the text (the transcripts).

The transcripts were coded using different codes for different content units.

5. Draw conclusions from the coded data.

After finishing coding, conclusions were made. The content units were classified into themes and sub-themes as illustrated in Table 3.1.

Table 3.1: Themes & sub-themes of content units

Themes	Sub-themes
Educational & Rating background	• Educational background • Rating background
Mental conditions	• Being busy lately • Returning from his last flight more than 24 hours • Feeling bored/exhausted/tired during rating

	• Any incident on the way to rating
Physical conditions	• Short ailments • Having a good sleep/rest the night before rating • Having enough rest/sleep the night before rating
Physical settings	• The room temperature was too warm or too cold or neither • The room was too dark or too lighted or neither • The room was too noisy • A preferred place to do the rating i.e. in an office, in a sound lab, or some places else
Rating strategies	• Listening without stopping strategy • Listening/stopping/note-taking strategy • Times of listening before rating • Note taking • Tape stopping (other than to take notes) • Stopping the tapes to listen for certain parts • Concentration on language or content or both • Focus on accuracy or fluency or both • Rating each criterion before or after the overall performance • Concentration on errors • Types of errors that raters listened for

	• Consideration on the relatedness/ relevance of the content as a factor in their ratings • Consideration on the quality of the content as a factor in their ratings • Consideration on the candidates‟ distinctive characteristics • Putting equal weight on all six criteria
Test tasks & Speech samples	• Degrees of test tasks • Duration of the speech samples • Appropriate duration of the speech samples • Rating three speech samples consecutively was too much • The maximum number of the speech samples that should be rated in one day
Interviewer/Interlocutor effects	• The interviewers/interlocutors tried to help/accommodate the candidate during the test • The interviewers/interlocutors attempted to simplify their speech to facilitate the candidates or to match the candidates‟ level of language • The interviewers performed their jobs appropriately
Candidates/Test-takers	• Taking the candidates‟ age into considerations

	• Taking the candidates‟ gender into considerations
Rating scales & descriptors	• Degrees of familiarity
	• Descriptor interpretation i.e. qualitatively or quantitatively
	• ICAO descriptor consultation before listening to the speech samples • ICAO descriptor consultation during listening to the speech samples • ICAO descriptor consultation after listening to the speech samples • Every English native speaker being at ICAO Level 6 • Being at ICAO Level 6 meaning equivalent to being an English native speaker
Cut-off score	• Awareness of Level 4 as the cut-off score • Consideration of the candidates‟ consequences as “pass” or “fail” in ratings
Personal relationship between raters and candidates	• Consideration of any personal relationship with the candidates
Scoring techniques	• Awareness of the overall score as the lowest among all six criteria

	• Consideration of score change after knowing that the overall score being based on the lowest score among all six criteria
Raters‟ harshness/leniency	• Self-consideration as being harsh, lenient or neither

As for coding, categories (verbal and non-verbal) were classified into meaning units and codes as shown in Table 3.2.

Table 3.2: Categories, meaning units & codes

Categories	Meaning units	Codes
Non-verbal	Sighs	Futility
	Pauses	Hesitation
	Long pauses	Stronger degree of hesitation
	Silence	Inability to explain
Verbal	“Yes”	Agreement
	“No”	Disagreement
	“Okay”	Approval/Assent/Acknowledgment

The other verbal expressions, which were exact words and straightforward, were not coded and categorized since they were manifest contents (see 2.11 for Content analysis).

After analyzing the data, the results and discussions are presented in Chapter 4.

CHAPTER IV
RESULTS AND DISCUSSIONS

This chapter presents the results and discussions of the research entitled "A study of Raters' Background Knowledge, Rater training, and Other Factors Affecting their Decision Making in Rating Thai Pilots' English Speaking Proficiency". The purposes of this study are:

- To investigate the effects of the different background knowledge of raters on their ratings of pilots‟ speaking ability.
- To explore the effects of rater training on their ratings of pilots‟ speaking ability.
- To examine the interaction effects between raters‟ background knowledge and their training on their ratings of pilots‟ speaking ability.
- To examine other factors affecting the decision-making of raters in their rating of pilots‟ speaking ability.

The study was conducted in order to test the hypotheses concerning the effects of raters‟ background knowledge and their training on rating pilots‟ English speaking proficiency as follows:

$H^1 1$: The linguistic raters will rate test takers‟ performance significantly and differently from operational raters ($p \leq .05$).

$H^1 2$: The raters who are trained in any rater training course will rate significantly and differently from those who are not ($p \leq .05$).

$H^1 3$: There are significant effects among types of raters, rater training and rating performance ($p \leq .05$).

The data were presented in tables and the interpretations of the tables were done in prose. The data were presented and discussed in three sections as follows:

Section One: Results and discussions about raters‟ rating scores

Section Two: Results and discussions obtained from the questionnaire

Section Three: Results and discussions concerning the factors affecting the raters‟ decision-making

These abbreviations will be used, OT = Operational/trained raters, OU= Operational/untrained raters, LT = Linguistic/trained raters, LU = Linguistic/untrained raters.

Section One: Results and discussions about raters" rating scores

Table 4.1 shows the scores each rater awarded to the speech sample number 1 in each criterion and the overall scores.

Table 4.1: Rating results for Speech sample no. 1 among four groups of raters

	Pronun-ciation	Structure	Vocabulary	Fluency	Compre-hension	Interactions	Overall
OT1	4	4	3	4	5	4	**4**
OT2	5	4	4	5	5	5	**4**
OT3	4	4	4	4	4	4	**4**
OT4	4	4	4	5	5	5	**4**
OT5	4	4	4	5	4	5	**4**
OU1	5	5	5	5	5	6	**5**
OU2	5	4	5	5	5	4	**5**
OU3	4	3	3	4	5	4	**4**
OU4	5	4	5	4	5	4	**4.5**
OU5	4	4	3	4	4	4	**4**
LT1	5	5	5	5	5	6	**5**
LT2	4	4	4	4	4	4	**4**
LT3	4	4	4	4	5	5	**4**
LT4	4	4	4	5	5	5	**4**
LT5	5	5	6	6	6	6	**5**
LU1	4	5	5	4	5	5	**5**
LU2	3	4	4	3	4	4	**4**
LU3	4	4	4	5	4	5	**4+**
LU4	5	5	5	5	5	5	**5**
LU5	4	4	4	3	4	4	**4**

Table 4.2 shows the scores each rater awarded to the speech sample number 2 in each criterion and the overall scores.

Table 4.2: Rating results for Speech sample no. 2 among four groups of raters

	Pronun-ciation	Structure	Vocabulary	Fluency	Compre-hension	Interactions	Overall
OT1	3	3	3	3	4	4	**3**
OT2	4	4	4	4	3	4	**3**
OT3	3	3	4	3	3	3	**3**
OT4	4	4	4	4	4	4	**4**
OT5	3	4	4	4	4	4	**3**
OU1	5	4	4	5	5	5	**5**
OU2	5	4	4	4	3	4	**4**
OU3	2	3	3	2	3	2	**3**
OU4	4	3	3	4	3	3	**3.3**
OU5	3	3	3	2	3	3	**3**
LT1	4	4	4	4	4	4	**4**
LT2	3	3	3	3	3	3	**3**
LT3	3	3	3	4	4	4	**3**
LT4	4	4	4	4	4	5	**4**
LT5	4	3	3	4	3	4	**3**
LU1	3	4	4	4	3	4	**4**
LU2	3	4	4	4	4	4	**4**
LU3	3	3	2	2	3	2	**3**
LU4	4	4	4	4	4	4	**4**
LU5	3	3	3	3	3	4	**3.5**

Table 4.3 shows the scores each rater awarded to the speech sample number 3 in each criterion and the overall scores.

Table 4.3: Rating results for Speech sample no. 3 among four groups of raters

	Pronun-ciation	Structure	Vocabulary	Fluency	Compre-hension	Interactions	Overall
OT1	4	3	3	4	4	4	**3**
OT2	4	3	3	4	4	4	**3**
OT3	5	4	4	4	4	5	**4**
OT4	4	3	4	4	4	4	**3**
OT5	4	4	4	4	5	4	**4**
OU1	4	4	4	5	4	4	**4**
OU2	5	4	4	4	4	4	**4**
OU3	4	2	2	3	3	3	**3**
OU4	5	4	5	5	4	5	**4.75**
OU5	4	4	3	3	3	4	**4**
LT1	5	3	3	4	4	4	**3**
LT2	4	3	3	3	4	4	**3**
LT3	3	3	3	3	3	3	**3**
LT4	3	3	3	3	3	4	**3**
LT5	4	4	4	4	5	5	**4**
LU1	5	5	5	5	5	5	**5**
LU2	3	4	3	3	4	4	**4**
LU3	4	3	3	3	4	4	**4**
LU4	4	4	4	4	4	4	**4**
LU5	4	4	4	4	3	3	**4**

The non-full digit scores provided by some untrained raters i.e. 4.5 and 4+ in Table 4.1, 3.3 and 3.5 in Table 4.2, and 4.75 in Table 4.3 are the raw scores that those raters considered the test-takers deserved even though it did not comply with the full digit score that ICAO requires. This is because they were not aware of and were not briefed about this requirement before rating.

According to the independent variables of the study (raters‟ background and raters‟ training), the 20 subjects were primarily categorized into two main groups, which were rater background and rater training. Each variable had two levels: rater background (operational/linguistic raters), rater training (trained/untrained raters). Figure 4.1 illustrates four groups of raters participating in the study and presents the sample means for ANOVA.

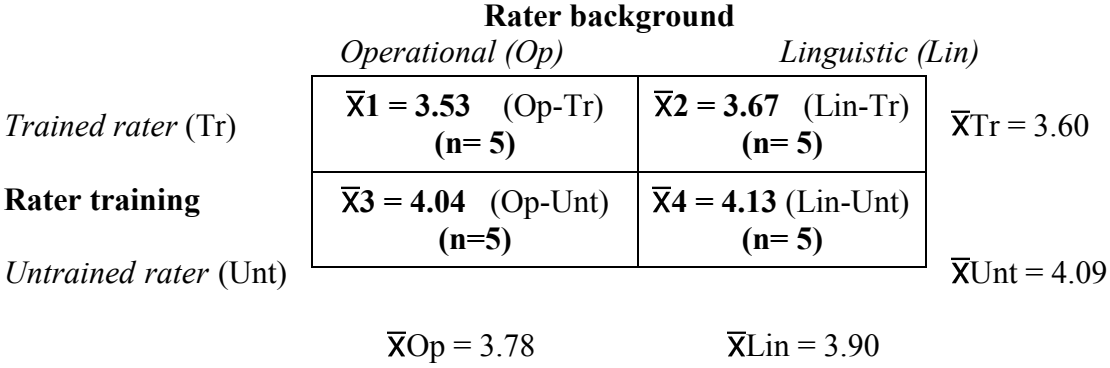

Figure 4.1: Sample Means for ANOVA

With regard to Figure 4.1, there are four groups of raters based on the two main variables. The subject groups include the group of operational/trained raters, operational/untrained raters, linguistic/trained raters, and linguistic/untrained raters. In each cell, there are five subjects assigned. All 20 subjects were required to rate three speech samples based on holistic scales of six levels provided by ICAO. Table 4.4 shows the descriptive statistics of raters rating the three speech samples.

Table 4.4: Descriptive Statistics of Raters

Speech samples	Raters Group 1 (Op/Tr)	Raters Group 2 (Op/Unt)	Raters Group 3 (Lin/Tr)	Raters Group 4 (Lin/Unt)	$\bar{X}$
1	$\bar{X}1 = 4.00$ SD. = 0 Max = 4 Min = 4	$\bar{X}2 = 4.50$ SD. = 0.5 Max = 5 Min = 4	$\bar{X}3 = 4.4$ SD. = 0.55 Max = 5 Min = 4	$\bar{X}4 = 4.50$ SD. = 0.5 Max = 5 Min = 4	$\bar{\bar{X}}1234 = 4.35$
2	$\bar{X}1 = 3.20$ SD. = 0.45	$\bar{X}2 = 3.66$ SD. = 0.85	$\bar{X}3 = 3.40$ SD. = 0	$\bar{X}4 = 3.70$ SD. = 0	$\bar{\bar{X}}1234 = 3.49$

	Max = 4 Min = 3	Max = 5 Min = 3	Max = 4 Min = 4	Max = 4 Min = 4	
3	$\bar{X}1 = 3.40$ SD. = 0 Max = 4 Min = 4	$\bar{X}2 = 3.95$ SD. = 0 Max = 4 Min = 4	$\bar{X}3 = 3.20$ SD. = 0 Max = 4 Min = 4	$\bar{X}4 = 4.20$ SD. = 0 Max = 4 Min = 4	$\bar{\bar{X}}1234 = 3.69$
$\bar{\bar{X}}$	$\bar{\bar{X}}1 = 3.53$	$\bar{\bar{X}}2 = 4.04$	$\bar{\bar{X}}3 = 3.67$	$\bar{\bar{X}}4 = 4.13$	

According to Table 4.4, and also Figure 4.1, the data show that the mean scores of the judgment for both groups of untrained raters are higher than those for trained raters, and the average scores rated from the linguistic raters are higher than those from the operational raters.

In order to test the three hypotheses mentioned earlier and answer the first, second and third research questions, the analysis of ANOVA was conducted. Regarding the judgment of each speech sample, the main and interaction effects of raters‟ background (operational and linguistic raters) and rater training (trained and untrained raters) are presented in the following ANOVA tables (4.5, 4.6, and 4.7).

Table 4.5: ANOVA summary table of “speech sample no.1”

Source	**SS**	***df***	**Mean Square**	**F**	**Sig.**
BACKGROUND	.113	1	.113	.529	.477
TRAINING	.312	1	.312	1.47	.243
BACKGROUND * TRAINING	.313	1	.313	1.47	.243
Error	3.40	16	.212		
Total	378.250	20			
Corrected Total	4.137	19			

$p \leq .05$

Bachman (2004) stated that if the F-ratio of the estimates between groups and within-group variances is larger than the F-critical, it means that the overall difference among the groups is not due to chance. Regarding Table 4.5 which focuses on the effects of variables on raters‟ decision-making for rating the first speech sample, when the observed values of F were compared with their critical values (F = 4.38, for the .05 level), it was found that all values of F calculated are less than the critical value [F (1, 16) = 0.53, 1.47, and 1.47, p > .05]. Therefore, the hypothesis studying the rating of the first speech sample was rejected. It might be concluded that the effect of different rater background (operational/linguistic raters), the effect of different rater training (trained/untrained raters), and the interaction effect between rater background and their training were not significantly different. In other words, the studied variables do not have any effect on the raters‟ decision-making.

Table 4.6: ANOVA summary table of "speech sample no.2"

Source	SS	*df*	Mean Square	F	Sig.
BACKGROUND	.072	1	.072	0.20	.659
TRAINING	.722	1	.722	2.02	.174
BACKGROUND * TRAINING	.032	1	.032	0.09	.768
Error	5.71	16	.357		
Total	250.140	20			
Corrected Total	6.54	19			

p ≤.05

Table 4.6 presents the results of the effects of variables on raters‟ decision-making for rating the second speech sample. The results correspond with the rating for the first speech sample in that all hypotheses fail to accept since the F-calculated was 0.20, 2.02, and 0.09 which were less than the F-critical (4.38) for the .05 level. Therefore, it might be said that, for rating the second speech sample, significant differences were not found in the effect of different rater background, the effect of different rater training, and the interaction effect between rater background and training, on decision-making in rating speaking proficiency.

Table 4.7: ANOVA summary table of "speech sample no.3"

Source	SS	*df*	Mean Square	F	Sig.
BACKGROUND	.003	1	.003	0.01	.916
TRAINING	3.003	1	3.003	11.05*	.004
BACKGROUND * TRAINING	.253	1	.253	0.93	.349
Error	4.35	16	.272		
Total	279.562	20			
Corrected Total	7.609	19			

*$p \leq .05$

Table 4.7 presents the main and interaction effects of rater background and rater training on raters" decision-making in rating speaking ability of "the speech sample number three". It was found that raters with different background have no effect on their decision-making in rating the speech sample. Correspondingly, there was no interaction effect between rater background and training on rating the speech sample. The evidence supporting these two cases is that the F-calculated (0.01 and 0.93) were less than the F-critical value (4.38). However, focusing on the factor of training, the F-calculated was 11.05 which was larger than the F-critical (4.38) for the .05 level. Therefore, it could be concluded that the difference in rater training (trained/untrained raters) had a significant effect on the decision-making of raters rating the speech sample number three, $F (1, 16) = 11.05$, $p < .05$. For better understanding of further explanation, the best way to interpret is to plot the means of the groups. The figure can make it easier for readers to understand what has happened between the levels of the factors (Hatch & Farhady, 1982). Figure 4.2 yields the result illustrating the training of raters significantly affects their decision-making in rating English proficiency of the speech sample no.3.

Figure 4.2: The effect of rater training on raters‟ decision-making in rating the speech sample no.3

The finding shows that the raters‟ decision-making in rating the speech sample number three is able to distinguish between trained and untrained raters. It can be seen from Figure 4.2 that the raters who had not been trained seemed to be more lenient in rating the speech sample no.3 than the raters who had been trained. The mean of the former is 4.08, while that of the latter is 3.30.

Regarding the first hypothesis testing of the speech samples, the main effect of the factor of raters‟ background on their decision-making in rating Thai pilots‟ English speaking proficiency, the results from Table 4.5, 4.6, and 4.7 show that the F-calculated (0.53, 0.20, and 0.01) were less than the F-critical (4.38) for the .05 level. Therefore, the hypothesis stated was rejected and it could be concluded that the difference in rater background (operational/linguistic raters) did not have a significant effect on raters‟ decision-making in their judgment. It might be said that operational raters and linguistic raters performed in rating English proficiency similarly. The mean of the first group was 3.78, while the mean of the second one was 3.90 (see Figure 4.1).

Concerning the second hypothesis, the main effect of raters‟ training on their judgment, the results from the two out of three speech samples (see Table 4.5 and 4.6) present that the F-calculated of this variable was 1.47 and 2.02 which were less than the F-critical (4.38) for the .05 level. Thus, the hypotheses stated for these two speech samples were rejected. However, considering the third speech sample (see Table 4.7), the F-calculated (11.05) was larger than the F-critical (4.38) which means that the difference in rater training (trained/untrained) affected raters‟ decision-making in their ratings of the third speech sample. So, the hypothesis to test this speech sample was accepted. Although it might not be consensually concluded that rater training has an effect on raters‟ judgment in speaking proficiency, it might be inferred that the training of raters affects more on their ratings than the factor of rater background. This can be seen when examining the F-value of the Sum of Squares (SS) of the training factor in that its value was more than the value of the factor of background. It means that when changing the value of the training while focusing on the same group of the rater background, it affected the output (raters‟ decision-making in rating) more than the change of the factor of background.

With regard to the interaction effect between the rater background and rater training, the results show that the significant interaction effect was not found. The third hypothesis stated was thus rejected. Figure 4.3 yields the results illustrating that both background factor and training factor do not significantly affect raters‟ decision.

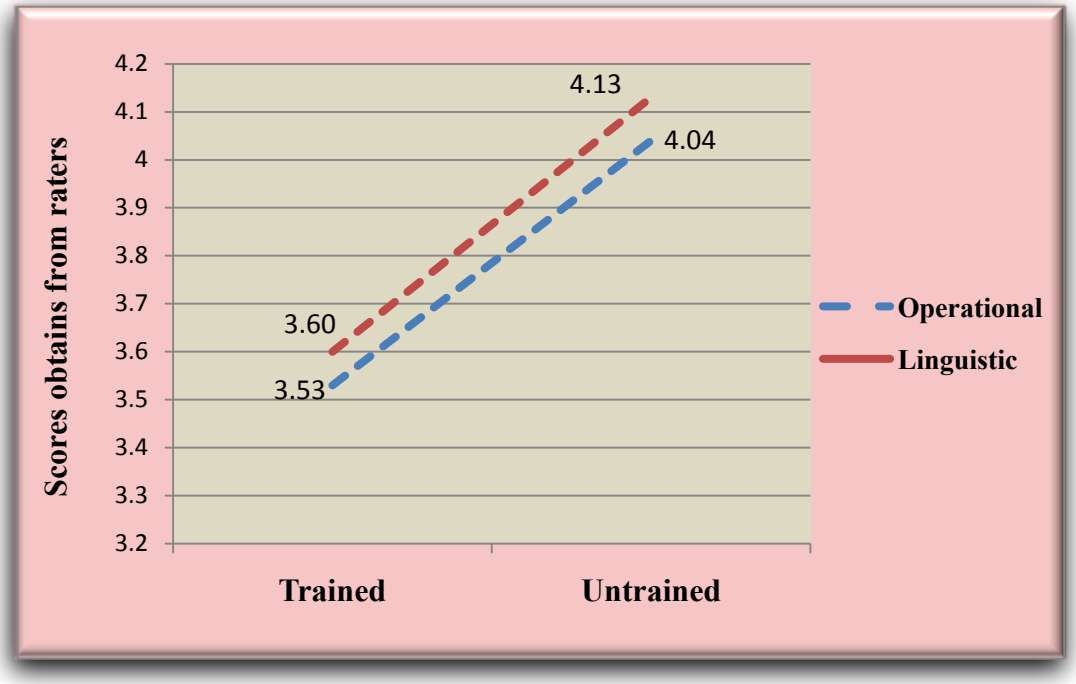

Figure 4.3: The interaction effect between rater background and rater training on raters‟ decision making in rating Thai pilots‟ English speaking proficiency

Figure 4.3 indicates that there was no significant interaction effect between the two independent variables on raters‟ judgment. However, from the graph it can be said that the untrained operational raters tended to rate in higher scores than the group of trained operational raters. The mean of the former is 4.04 while that of the latter is 3.53. Similarly, the untrained linguistic raters judged the samples‟ speaking proficiency with higher scores than the trained linguistic raters. The means are 4.13 and 3.60 respectively. It might be concluded that both rater background groups seemed to be more lenient in rating speaking proficiency when they are untrained.

In conclusion, to answer the first three research questions, the hypotheses were tested and it was found that all hypotheses were rejected meaning that both rater background and rater training did not significantly affect raters‟ decision-making in rating Thai pilots‟ English speaking proficiency, in both main and interaction effects. However, the factor of training seemed to significantly affect more than the factor of background on the dependent variable.

In order to confirm the findings and gain more supporting information, the four groups of raters‟ rated scores, based on each individual and the overall criterion, were compared in Table 4.8 to Table 4.14.

Table 4.8: ANOVA Table comparing rating of "pronunciation" criterion among four groups of raters

Pronunciation					
Speech sample 1					
Source	**SS**	***df***	**MS**	**F**	**Sig.**
Between Groups	1	3	0.33	1.02	0.41
Within Groups	5.2	16	0.33		
Speech sample 2					
Source	**SS**	***df***	**MS**	**F**	**Sig.**
Between Groups	1	3	0.33	0.54	0.67
Within Groups	10	16	0.63		
Speech sample 3					
Source	**SS**	***df***	**MS**	**F**	**Sig.**
Between Groups	1	3	0.33	0.78	0.52
Within Groups	6.8	16	0.43		

*p≤.05

According to the criterion of "pronunciation" shown in Table 4.8, it was found that each F-calculated value (1.02, 0.54 and 0.78) was less than the F-critical value (3.34) for the .05 level. Therefore, it might be said that there was no significant difference among four groups of raters‟ judgment in rating each speech sample in terms of their pronunciation.

Table 4.9: ANOVA Table comparing rating of "structure" criterion among four groups of raters

Structure					
Speech sample 1					
Source	**SS**	***df***	**MS**	**F**	**Sig.**
Between Groups	0.8	3	0.27	0.97	0.43
Within Groups	4.4	16	0.27		
Speech sample 2					
Source	**SS**	***df***	**MS**	**F**	**Sig.**
Between Groups	0.2	3	0.07	0.22	0.88
Within Groups	4.8	16	0.3		
Speech sample 3					
Source	**SS**	***df***	**MS**	**F**	**Sig.**
Between Groups	1.75	3	0.58	1.29	0.31
Within Groups	7.2	16	0.45		

*p≤.05

Table 4.9 presents that the F-calculated values were 0.97, 0.22, and 1.29 which were less than the F-critical value (3.34) for the .05 level. It might be interpreted that there was no significant difference between four groups of raters in rating three speech samples considering the "structure" criterion.

Table 4.10: ANOVA Table comparing rating of "vocabulary" criterion among four groups of raters

Vocabulary					
Speech sample 1					
Source	**SS**	***df***	**MS**	**F**	**Sig.**
Between Groups	1.75	3	0.58	0.93	0.44
Within Groups	10	16	0.63		

Speech sample 2					
Source	**SS**	***df***	**MS**	**F**	**Sig.**
Between Groups	0.6	3	0.2	0.50	0.69
Within Groups	6.4	16	0.4		
Speech sample 3					
Source	**SS**	***df***	**MS**	**F**	**Sig.**
Between Groups	0.95	3	0.32	0.50	0.68
Within Groups	10	16	0.62		

*p≤.05

Table 4.10 might be interpreted that the differences among four groups of raters had no significant effect on the decision-making of raters rating all speech samples in regard to the criterion of "vocabulary", as the F-calculated values (0.93, 0.50, and 0.50) were less than the F-critical value (3.34) for the .05 level.

Table 4.11: ANOVA Table comparing rating of "fluency" criterion among four groups of raters

Fluency					
Speech sample 1					
Source	**SS**	***df***	**MS**	**F**	**Sig.**
Between Groups	1.75	3	0.58	1.01	0.41
Within Groups	9.2	16	0.58		
Speech sample 2					
Source	**SS**	***df***	**MS**	**F**	**Sig.**
Between Groups	0.55	3	0.18	0.24	0.87
Within Groups	12.4	16	0.77		

Speech sample 3					
Source	**SS**	***df***	**MS**	**F**	**Sig.**
Between Groups	1.2	3	0.4	0.80	0.51
Within Groups	8	16	0.5		

*p≤.05

Regarding the criterion of "fluency" presented in Table 4.11, it was found that each F-calculated value (1.01, 0.24 and 0.80) was less than the F-critical value (3.34) for the .05 level. Therefore, it might be said that there was no significant difference among four groups of raters" judgment in rating each speech sample according to this matter.

Table 4.12: ANOVA Table comparing rating of "comprehension" criterion among four groups of raters

Comprehension					
Speech sample 1					
Source	**SS**	***df***	**MS**	**F**	**Sig.**
Between Groups	1	3	0.33	1.03	0.41
Within Groups	5.2	16	0.33		
Speech sample 2					
Source	**SS**	***df***	**MS**	**F**	**Sig.**
Between Groups	0.2	3	0.06	0.16	0.92
Within Groups	6.8	16	0.43		
Speech sample 3					
Source	**SS**	***df***	**MS**	**F**	**Sig.**
Between Groups	1	3	0.3	0.78	0.52
Within Groups	6.8	16	0.43		

*p≤.05

Table 4.12 illustrates that the F-calculated values were 1.03, 0.16, and 0.78 which were less than the F-critical value (3.34) for the .05 level. It might be interpreted that there was no significant difference among four groups of raters in rating three speech samples considering the "comprehension" criterion.

Table 4.13: ANOVA Table comparing rating of "interactions" criterion among four groups of raters

Interactions					
Speech sample 1					
Source	**SS**	***df***	**MS**	**F**	**Sig.**
Between Groups	1.8	3	0.6	1.14	0.37
Within Groups	8.4	16	0.53		
Speech sample 2					
Source	**SS**	***df***	**MS**	**F**	**Sig.**
Between Groups	1	3	0.33	0.48	0.70
Within Groups	11.2	16	0.7		
Speech sample 3					
Source	**SS**	***df***	**MS**	**F**	**Sig.**
Between Groups	0.15	3	0.05	0.12	0.95
Within Groups	6.8	16	0.43		

*p≤.05

Table 4.13 presents that the difference among four groups of raters had no significant difference in the decision-making of raters rating the all speech sample according to the criterion of "interactions", as the F-calculated values (1.14, 0.48, and 0.12) were less than the F-critical value (3.34) for the .05 level.

Table 4.14: ANOVA Table comparing rating of "overall" criterion among four groups of raters

Overall					
Speech sample 1					
Source	**SS**	***df***	**MS**	**F**	**Sig.**
Between Groups	0.85	3	0.28	1.42	0.27
Within Groups	3.2	16	0.2		
Speech sample 2					
Source	**SS**	***df***	**MS**	**F**	**Sig.**
Between Groups	0.83	3	0.28	0.77	0.53
Within Groups	5.71	16	0.36		
Speech sample 3					
Source	**SS**	***df***	**MS**	**F**	**Sig.**
Between Groups	3.26	3	1.08	3.99*	0.03
Within Groups	4.35	16	0.27		

*$p \leq .05$

Table 4.14 illustrates that the F-calculated values were 1.42, 0.77, and 3.99. F-calculated values of the first two speech samples were less than the F-critical (3.34); on the other hand, the F-calculated value of the third speech sample was larger than the F-critical (3.34) for the .05 level. It might be interpreted that there was no significant difference among four groups of raters in rating the speech samples 1 and 2 considering "overall" criterion. However, there was a significant difference among four groups of raters in rating the speech sample 3.

In conclusion, according to Table 4.8 to Table 4.14, the finding indicates that, except the criterion of "overall", there was no significant difference among four groups of raters in rating three speech samples considering each criterion. Focusing on the rating of the "overall" criterion in the speech sample 3, the F-calculated value was 3.99 which was larger

than the F-critical (3.34) for the .05 level. Although it seems to be that the difference in raters‟ background (linguistic or operational raters) and their training (trained or untrained raters) had a significant effect in raters‟ rating the “overall” criterion, $F(3,16) = 3.99$, $p \leq .05$, it might not be concluded that, for all criteria, these two factors affect their decision-making in rating Thai pilots' English speaking proficiency. Therefore, it supports the finding mentioned in the first section that both rater background and rater training did not significantly affect raters‟ decision-making in rating Thai pilots‟ English speaking proficiency, in both main and interaction effects.

Section Two: Results and discussions obtained from the questionnaire

In this section, the results obtained from the questionnaire are shown according to the raters‟ answers starting from the linguistic/trained (LT), linguistic/untrained (LU), operational/trained (OT), and operational/untrained (OU) groups.

The data obtained from the linguistic/trained rater questionnaires revealed that all of them were female. Three of them aged between 31 to 40 years while the other two were between 51 to 60 years old. Three raters graduated with master‟s degrees. Among these, two raters were Ph.D. students. The remaining two hold bachelor‟s degrees. All of the linguistic/trained raters were language teachers. Four had experience in the occupation between 11 to 15 years. Only one had less experience, which was between 6 to 10 years. All of their first language was Thai. They studied English for more than 16 years. Two of them considered their English proficiency level as „very good‟ while the other three as „good‟. Everyone had formal rater training. Four raters had passed two rater training courses, namely TOEIC language proficiency interviewing/rating training course in 2006 and TRAINAIR Standardized Training Packages (STPs) interviewer/rater training course organized by Civil Aviation Training Center (CATC) in 2007. Only one rater was trained once in TRAINAIR Standardized Training Packages (STPs) interviewer/rater training course organized by Civil Aviation Training Center (CATC) in 2007.

Two raters accepted that the level of their exposure to various English native speakers‟ accents were „very much‟ while the other three were „much‟. Their degrees of

exposure to various Asian English speakers" accents were „very much" (1), „much" (3) and „some" (1). The linguistic/trained raters seemed to be less familiar with European English accents since three of them answered „some" and the other three answered „much". It might be because they were all language teachers, and their familiarity with linguistic terms were „very much" (2) and „much" (3). Since all of them were English language teachers working for Thai Airways Flight Crew Language Training Department, they were quite familiar with aviation operations and aeronautical communication in the level of „very much" (1), „much" (2) and „some" (2). All of them had experience in language assessment in the level of „very much" (2) and „much" (3) which are the same level for their familiarity with using language descriptors. They were also familiar with ICAO language proficiency scale in the level of „very much" (3) and „much" (2).

Four linguistic/trained raters stated that they „sometimes" consulted the details of each ICAO descriptor before listening to the speech samples. Only one „frequently" did it. Two raters „frequently" checked the details during listening. One „sometimes" did it, one „always" and one „rarely" did it during listening. After listening, two raters „frequently" turned to the details of ICAO descriptors. The other two „always" checked it and one „sometimes" did it after listening. Two raters „sometimes" listened to the speech samples before giving their final scores. One „always", one „frequently" and one „rarely" did it. Three raters „always" took notes while rating. The other two did it „frequently". No rater in this group „never" took notes at all. Three raters „frequently" stopped the tapes for a reason while rating. The other two did it „sometimes". Two raters in this group „always" stopped to listen for certain parts of the samples. Two „frequently" did it while another „sometimes" stopped it. Three raters accepted that they „always" concentrated on the errors made by the candidates. The other two did it „frequently". Two raters answered that they „frequently" considered the relevance of the content as a factor in their ratings while the other three „always" did it. No rater said that she „never" considered it as a factor in her rating.

All linguistic/trained raters said that they had been busy recently before rating. One admitted that she felt bored/exhausted/tired during rating. The others said „no". One rater in this group said that she had some kind of short-term ailments but none had long-term ones.

One of them remarked that she did not sleep well and did not get enough rest before rating. No one complained about the setting as being too cold, too warm, too dark, too lighted, or too noisy. No rater listened to the speech samples from the beginning to the end without stopping at least once. All raters confirmed that they weighted each criterion equally before giving the final score. Two of them also accepted that they considered the quality of the content the candidates gave as a factor in their ratings. Three did not. One rater thought that the test tasks were easy while one thought that the test tasks were difficult. Three raters thought that the speech samples were too short while no one thought that they were too long. Two raters felt that rating three speech samples consecutively was too much while the others did not. All raters thought that the interlocutors performed their jobs appropriately, though three raters thought that they tried to help/accommodate the candidates and two raters thought that they tried to simplify the speech to facilitate the candidates.

Only one linguistic/trained rater admitted that she considered the candidates‟ age in her ratings. The same rater also accepted that she considered the candidates‟ overall attitudes but she did not consider their genders. Three raters felt that the candidates were nervous during testing but only one of them said that she sympathized for it in her ratings. The majority of the raters (four) declared that they did not compare a candidate with other candidates. Only one did it. One rater thought that English native speakers must also be at ICAO Level 6 while another rater thought that being at ICAO Level 6 was equivalent to being an English native speaker. All raters said that they knew about the ICAO-required „cut-off‟ score. Only one of them said that she did not consider the consequences of the candidates as pass/fail in her ratings while the other four said they did. One rater considered changing the scores she already gave. Lastly, only one rater considered herself as lenient while another one considered herself as a harsh rater.

The summary of the linguistic/trained raters‟ answers to the questionnaire is shown in Table 4.15 below:

Table 4.15: Linguistic/Trained Raters‟ answers to the questionnaire

	LT1	LT2	LT3	LT4	LT5
1. Gender	Female	Female	Female	Female	Female
2. Age (Years)	51-60	31-40	31-40	51-60	31-40
3. Educational level	B.A. (English)	M.A. (Linguistics) Ph.D. candidate (Linguistics)	M.S. (Education) M.A. (Teaching English as a foreign language)	B.A. (English teaching)	M.A. (Linguistics) Ph.D. student (Higher education)
4. Occupation	Language teacher	Language teacher	Language teacher	Language teacher	Language teacher
5. Years of being in the occupation	11-15	11-15	6-10	11-15	11-15
6. First language(L1)	Thai	Thai	Thai	Thai	Thai
7. Duration of English study (years)	> 16 years	> 16 years	> 16 years	> 16 years	> 16 years
8. Level of English proficiency	Good	Good	Good	Very good	Very good
9. Formal rater training and the course name(s)	Yes *	Yes *	Yes *	Yes *	Yes **
10. Exposure to various English native speakers‟ accents	Very much	Very much	Much	Much	Much
11. Exposure to Asian English accents	Very much	Much	Much	Some	Much
12. Exposure to European English accents	Much	Much	Some	Some	Much
13. Degree of familiarity with linguistic terms	Very much	Very much	Much	Much	Much
14. Degree of familiarity with aviation operations and aeronautical communication	Very much	Much	Some	Some	Much
15. Experience in language assessment	Very much	Very much	Much	Much	Much
16. Familiarity with using language descriptors	Very much	Very much	Much	Much	Much
17. Familiarity with ICAO	Very much	Very much	Much	Very much	Much

language proficiency rating scale					
18. Frequency of consulting the details of each ICAO descriptor in Doc. 9835 <u>before</u> listening to the speech samples	Frequently	Sometimes	Sometimes	Sometimes	Sometimes
19. Frequency of consulting the details of each ICAO descriptor in Doc. 9835 <u>during</u> listening to the speech samples	Always	Frequently	Frequently	Rarely	Sometimes
20. Frequency of consulting the details of each ICAO descriptor in Doc. 9835 <u>after</u> listening to the speech samples	Always	Frequently	Always	Sometimes	Frequently
21. Frequency of listening to the given speech samples <u>before</u> giving the final score	Frequently	Sometimes	Rarely	Sometimes	Always
22. Frequency of taking notes <u>while</u> rating	Always	Always	Frequently	Always	Frequently
23. Frequency of stopping the tapes for any reason <u>while</u> rating	Frequently	Sometimes	Sometimes	Frequently	Frequently
24. Frequency of stopping to listen for certain parts from the speech samples	Always	Sometimes	Frequently	Frequently	Always
25. Frequency of concentrating on errors made by the speaker	Always	Always	Always	Frequently	Frequently
26. Frequency of considering the relatedness/rele	Frequently	Always	Always	Frequently	Always

vance of the content as a factor in your rating					
27. Having been busy lately?	Yes	Yes	Yes	Yes	Yes
28. Feeling bored/ exhausted/ tired during rating?	No	No	No	No	Yes
29. Having any short-term ailments?	No	No	Yes	No	No
30. Having any long-term ailments?	No	No	No	No	No
31. Having a good sleep/rest last night?	Yes	Yes	Yes	Yes	No
32. Had enough sleep/rest?	Yes	Yes	Yes	Yes	No
33. Was the room too cold?	No	No	No	No	No
34. Was the room too warm?	No	No	No	No	No
35. Was the room too dark?	No	No	No	No	No
36. Was the room too lighted?	No	No	No	No	No
37. Was the room too noisy?	No	No	No	No	No
38. Listening to the given speech sample from the beginning to the end without stopping at least once before rating?	No	No	No	No	No
39. Weighting each criterion equally before giving the final score?	Yes	Yes	Yes	Yes	Yes
40. Considering the quality of the content the candidates give as a factor in rating?	No	No	Yes	Yes	Yes
41. The test tasks were easy?	Yes	No	No	No	No
42. The test tasks were difficult?	No	No	No	Yes	No
43. The speech samples were	No	No	Yes	Yes	Yes

too short?					
44. The speech samples were too long?	No	No	No	No	No
45. Rating three speech samples consecutively was too much?	Yes	No	No	No	Yes
46. The interviewers/ interlocutors tried to help/ accommodate the candidate during the test?	No	Yes	No	Yes	Yes
47. The interviewers/ interlocutors performed their jobs appropriately/effectively as they should have?	Yes	Yes	Yes	Yes	Yes
48. The interviewers/ interlocutors attempted to simplify their speech to facilitate the candidates or to match the candidates‟ level of language?	No	Yes	No	No	Yes
49. Considering the candidates‟ age in rating?	No	No	No	No	Yes
50. Considering the candidates‟ gender in rating?	No	No	No	No	No
51. Considering the global/overall attitudes of the candidates?	No	No	No	No	Yes
52. The candidates were nervous during the test?	No	Yes	No	Yes	Yes
53. Sympathize for that nervousness in rating?	No	No	No	No	Yes
54. Comparing the candidate with	No	No	No	No	Yes

other candidates in rating?					
55. Every English native speaker must also be ICAO Level 6?	No	Yes	No	No	No
56. Being ICAO Level 6 equivalent to being an English native speaker?	No	No	No	No	Yes
57. Knowing that the „cut-off" score for this ICAO assessment is level 4?	Yes	Yes	Yes	Yes	Yes
58. Considering the consequence of the candidates as being passed or fail in rating?	No	Yes	Yes	Yes	Yes
59. Considering changing the scores already gave them?	No	No	No	No	Yes
60. Considering as being a lenient rater	No	No	No	No	Yes
61. Considering as being a harsh rater	No	No	Yes	No	No

* TOEIC language proficiency interviewer/rater training course in 2006 & TRAINAIR Standardized Training Packages (STPs) interviewer/rater training course organized by Civil Aviation Training Center (CATC) in 2007.
** TRAINAIR Standardized Training Packages (STPs) interviewer/rater training course organized by Civil Aviation Training Center (CATC) in 2007.

The data received from the linguistic/untrained rater questionnaires unveiled that the group of the linguistic/untrained raters was a combination of three females and two males. All of them aged between 31 to 40 years old. This whole group graduated with master"s degrees. Furthermore, every rater was Ph.D. candidate, three in the field of English teaching instruction & curriculum development and two in language assessment & evaluation. Four linguistic/untrained raters were language teachers while the remaining one was a linguist. Two of them had experience in the occupation between 6 to 10 years. The other three had

less experience, which was between 1 to 5 years. All of them spoke Thai as their first language. Almost all of them studied English for more than 16 years. Only one answered that she studied English between 11 to 15 years. Two of them considered their English proficiency level as „very good" while the other two as „good". One of the linguistic/untrained rater interestingly considered his English proficiency as „native-like/near native". No one had previous formal rater training, even those two who majored in language assessment & evaluation.

Three of the raters in this batch quoted their levels of exposure to various English native speakers" accents as „much" while the other two as „some". Their degrees of exposure to Asian English accents were varied as two „some", one „much" and two „very much". They seemed not to be so familiar with the European English accents since two of them answered their degrees as „little" while the other three as „some". Because they were language teachers or linguists, their consideration of the degrees of familiarity with linguistic terms were „much" (3) and „some" (2). On the contrary, four linguistic/untrained raters referred to their familiarity with aviation operations and aeronautical communication as „little". Only one rater considered it as „some". When being asked about their experience in language assessment, a rater"s answer was „very much", two answered as „much" and the other two as „some". Their degrees of familiarity with using language descriptors were: one „very much", three „much" and one „some". Three raters in this group were unfamiliar with the ICAO language proficiency rating scale by answering „none" while the other two answered „little".

In spite of their unfamiliarity with the ICAO language proficiency rating scale, two linguistic/untrained raters answered that they „rarely" consulted the details of each ICAO descriptor before listening to the speech samples. Two did it „sometimes" and only one „always" did it. Two raters „sometimes" checked the details during listening. On one hand, the same rater, who „always" checked it before listening, also „always" did it during and after listening. On the other hand, the same raters who „rarely" looked at it before listening, „never" consulted the details both during and after listening. After listening, two raters „always" turned to the details of ICAO descriptors. One „frequently" checked it and another „did it „sometimes". Three raters „sometimes" listened to the speech samples before giving their

final scores. The other two „frequently“ did it. Three raters „frequently“ took notes while rating. One „always“ did it and one „sometimes“ did it. No rater in this group took notes at all. Two raters „rarely“ stopped the tapes for a reason while rating. The other two did it „sometimes“ and one „frequently“ stopped it. Two raters in this group „sometimes“ stopped to listen for certain parts of the samples. The other two „rarely“ did it while another „frequently“ stopped it. Three raters accepted that they „sometimes“ concentrated on the errors made by the candidates. One did it „always“ and another „rarely“. Three raters answered that they „always“ considered the relevance of the content as a factor in their ratings while the other two did it „sometimes“.

All linguistic/untrained raters said that they had been busy lately before rating. Two raters admitted that they felt bored/exhausted/tired during rating. The others said „no“. No rater in this group had any kind of short-term and long-term ailments. Everyone said that s/he had well and enough rest before rating. No complaint was made about the setting as being too cold, too warm, too dark, too lighted, or too noisy from this group of raters. Only one rater listened to the speech samples from the beginning to the end without stopping at least once while the others did not. Four raters admitted that they weighted each criterion equally before giving the final score. However, one rater admitted that she did not weight each criterion equally. All of them accepted that they considered the quality of the content the candidates gave as a factor in their ratings. None of them thought that the test tasks were easy and three raters thought that the test tasks were difficult. No rater thought that the speech samples were too short while one thought that they were too long. Four raters did not feel that rating three speech samples consecutively was too much while only one did. All raters thought that the interlocutors performed their jobs appropriately, though one rater thought that they tried to help/accommodate the candidates and all raters thought that the interlocutors tried to simplify the speech to facilitate the candidates.

No linguistic/untrained rater considered the candidates‘ age in their ratings but one rater considered both gender and overall attitudes in her ratings. All raters, except one, felt that the candidates were nervous during testing and two of them said that they sympathized for it in their ratings. A greater number of the raters (four) declared that they compared a

candidate with other candidates. Only one did not. All raters did not think that English native speakers must also be at ICAO Level 6 but one of them thought that being at ICAO Level 6 was equivalent to being an English native speaker. Three raters said that they knew about the ICAO-required „cut-off" score and just one of them accepted he considered the consequences of the candidates as pass/fail in his ratings. No rater considered changing the scores s/he already gave. Finally, almost all raters (four) considered themselves as lenient while only one considered himself as a harsh rater.

The summary of the linguistic/untrained raters" answers to the questionnaire is shown in Table 4.16 below:

Table 4.16: Linguistic/Untrained Raters" answers to the questionnaire

	LU1	LU2	LU3	LU4	LU5
1. Gender	Female	Female	Female	Male	Male
2. Age (Years)	31-40	31-40	31-40	31-40	31-40
3. Educational level	Ph.D. candidate (Language assessment & evaluation)	Ph.D. candidate (English teaching instruction & curriculum development)	Ph.D. candidate (English teaching instruction & curriculum development)	Ph.D. candidate (Language assessment & evaluation)	Ph.D. candidate (English teaching instruction & curriculum development)
4. Occupation	Language teacher	Language teacher	Linguist	Language teacher	Language teacher
5. Years of being in the occupation	1-5	6-10	1-5	1-5	6-10
6. First language(L1)	Thai	Thai	Thai	Thai	Thai
7. Duration of English study (years)	> 16	11-15	> 16	> 16	> 16
8. Level of English proficiency	Very good	Good	Good	Very good	Native-like/ Near native
9. Formal rater training and the course name(s)	No	No	No	No	No
10. Exposure to various English native speakers" accents	Much	Some	Some	Much	Much
11. Exposure to Asian English accents	Some	Much	Some	Very much	Very much

12. Exposure to European English accents	Little	Little	Some	Some	Some
13. Degree of familiarity with linguistic terms	Much	Some	Some	Much	Much
14. Degree of familiarity with aviation operations and aeronautical communication	Little	Little	Little	Some	Little
15. Experience in language assessment	Very much	Some	Some	Much	Much
16. Familiarity with using language descriptors	Very much	Much	Much	Much	Some
17. Familiarity with ICAO language proficiency rating scale	None	Little	None	None	Little
18. Frequency of consulting the details of each ICAO descriptor in Doc. 9835 <u>before</u> listening to the speech samples	Sometimes	Rarely	Sometimes	Always	Rarely
19. Frequency of consulting the details of each ICAO descriptor in Doc. 9835 <u>during</u> listening to the speech samples	Sometimes	Never	Sometimes	Always	Frequently
20. Frequency of consulting the details of each ICAO descriptor in Doc. 9835 <u>after</u> listening to the speech samples	Always	Never	Sometimes	Always	Frequently
21. Frequency of listening to the given speech	Sometimes	Sometimes	Sometimes	Frequently	Frequently

samples <u>before</u> giving the final score					
22. Frequency of taking notes <u>while</u> rating	Always	Frequently	Sometimes	Frequently	Frequently
23. Frequency of stopping the tapes for any reason <u>while</u> rating	Sometimes	Sometimes	Frequently	Rarely	Rarely
24. Frequency of stopping to listen for certain parts from the speech samples	Sometimes	Sometimes	Frequently	Rarely	Rarely
25. Frequency of concentrating on errors made by the speaker	Sometimes	Sometimes	Rarely	Always	Sometimes
26. Frequency of considering the relatedness/relevance of the content as a factor in your rating	Always	Sometimes	Sometimes	Always	Always
27. Having been busy lately?	Yes	Yes	Yes	Yes	Yes
28. Feeling bored/ exhausted/ tired during rating?	No	Yes	No	Yes	No
29. Having any short-term ailments?	No	No	No	No	No
30. Having any long-term ailments?	No	No	No	No	No
31. Having a good sleep/rest last night?	Yes	Yes	Yes	Yes	Yes
32. Had enough sleep/rest?	Yes	Yes	Yes	Yes	Yes
33. Was the room too cold?	No	No	No	No	No
34. Was the room too warm?	No	No	No	No	No
35. Was the room too dark?	No	No	No	No	No
36. Was the room too lighted?	No	No	No	No	No
37. Was the room	No	No	No	No	No

too noisy?					
38. Listening to the given speech sample from the beginning to the end without stopping at least once before rating?	No	No	No	No	Yes
39. Weighting each criterion equally before giving the final score?	Yes	No	Yes	Yes	Yes
40. Considering the quality of the content the candidates give as a factor in rating?	Yes	Yes	Yes	Yes	Yes
41. The test tasks were easy?	No	No	No	No	No
42. The test tasks were difficult?	No	Yes	Yes	No	Yes
43. The speech samples were too short?	No	No	No	No	No
44. The speech samples were too long?	No	No	No	No	Yes
45. Rating three speech samples consecutively was too much?	No	No	Yes	No	No
46. The interviewers/ interlocutors tried to help/ accommodate the candidate during the test?	No	Yes	No	No	No
47. The interviewers/ interlocutors performed their jobs appropriately/ effectively as they should have?	Yes	Yes	Yes	Yes	Yes
48. The interviewers/	Yes	Yes	Yes	Yes	Yes

interlocutors attempted to simplify their speech to facilitate the candidates or to match the candidates‟ level of language?					
49. Considering the candidates‟ age in rating?	No	No	No	No	No
50. Considering the candidates‟ gender in rating?	No	No	Yes	No	No
51. Considering the global/overall attitudes of the candidates?	No	No	Yes	No	No
52. The candidates were nervous during the test?	Yes	Yes	Yes	No	Yes
53. Sympathize for that nervousness in rating?	Yes	No	No	No	Yes
54. Comparing the candidate with other candidates in rating?	No	Yes	Yes	Yes	Yes
55. Every English native speaker must also be at ICAO Level 6?	No	No	No	No	No
56. Being at ICAO Level 6 equivalent to being an English native speaker?	Yes	No	No	No	No
57. Knowing that the „cut-off‟ score for this ICAO assessment is level 4?	No	No	No	Yes	Yes
58. Considering the consequence of the candidates	No	No	No	Yes	No

as pass or fail in rating?					
59. Considering changing the scores already given to them?	No	No	No	No	No
60. Considering as being a lenient rater	Yes	Yes	Yes	Yes	No
61. Considering as being a harsh rater	No	No	No	No	Yes

The data acquired from the operational/trained rater questionnaires disclosed that, since Thai Airways has no female pilot, all of the operational/trained raters were males. They aged between 31 to 40 years old. This whole group graduated with bachelor"s degrees in various fields, none in the field of language or language-related. Two of them had experience in the occupation between 11 to 15 years. The other two had less experience, which was between 6 to 10 years. Only one had the least experience that was between 1 to 5 years. This was the only rater group that three of them spoke English as their first language while the other two were Thai. One of the two Thai studied English for more than 16 years and considered his English proficiency as „very good". The other studied English between 11 to 15 years and considered his English proficiency as „good". All of them had formal rater training. One rater went through TOEIC language proficiency interviewing/rating training course in 2006 only. The other one was trained by TRAINAIR Standardized Training Packages (STPs) interviewer/rater training course organized by Civil Aviation Training Center (CATC) in 2007. The remaining three took both rater training courses.

Their levels of exposure to various English native speakers" accents were rather high as „very much" (4) and „much" (1). Their degrees of exposure to Asian English accents were also high as two „much" and three „very much". They seemed to be less familiar with the European English accents since two of them answered their degrees as „much" while the other two as „some". Only one considered his as „very much". Perhaps, because they had no educational background in language or linguistic, they modestly considered their degrees of familiarity with linguistic terms as „little" (1) and „some" (3). Only one rater answered this as „much". On the contrary, four raters in this batch referred to their familiarity with aviation

operations and aeronautical communication as „very much". Only one rater considered it as „much". When being asked about their experience in language assessment, two raters answered „very much", and the other three answered „some". Their degrees of familiarity with using language descriptors were two „much" and three „some". Two raters in this group put their familiarity with the ICAO language proficiency rating scale as „much" while the other two as „some". Only one rater considered his as „very much".

Three operational/trained raters answered that they „frequently" consulted the details of each ICAO descriptor before listening to the speech samples. One „always" did it and another „rarely" did it. Three raters „frequently" checked the details during listening. One „always" did it and the same guy who „rarely" did it before listening also „rarely" did it during listening. After listening, three raters „frequently" turned to the details of the ICAO descriptors. The other two „always" checked it after listening. Three raters „sometimes" listened to the speech samples before giving their final scores. The other two „always" did it. Three raters „always" took notes while rating. One did it „frequently" and one „sometimes" did it. No rater in this group took notes at all. Two raters „frequently" stopped the tapes for a reason while rating. The other two did it „sometimes" and one „rarely" stopped it. Three raters in this group „sometimes" stopped to listen for certain parts of the samples. One „rarely" did it while another „frequently" stopped it. Three raters accepted that they „frequently" concentrated on the errors made by the candidates. The other two did it „sometimes". Two raters answered that they „frequently" considered the relevance of the content as a factor in their ratings while the other two did it „sometimes". Only one rater said that he „never" considered it as a factor in his rating.

All operational/trained raters said that they had been busy recently before rating. Three raters admitted that they felt bored/exhausted/tired during rating. The others said „no". One rater in this group said that he had some kind of short ailments but none had long term ones. One of them remarked that he did not have well and enough rest before rating. One complained about the setting as being too cold but no one found it too warm, too dark, or too lighted. Only one rater said that the room was too noisy. Two raters did not listen to the speech samples from the beginning to the end without stopping at least once while the other

three did. Four raters admitted that they weighted each criterion equally before giving the final score. However, one rater admitted that he did not weight each criterion equally. Two of them also accepted that they considered the quality of the content the candidates gave as a factor in their ratings. Three did not. Two of them thought that the test tasks were easy. The others did not think so. None thought that the test tasks were difficult. One rater thought that the speech samples were too short while no one thought that they were too long. Four raters felt that rating three speech samples consecutively was too much while only one did not. All raters thought that the interlocutors performed their jobs appropriately, though two raters thought that they tried to help/accommodate the candidates and three raters thought that they tried to simplify the speech to facilitate the candidates.

No operational/trained rater considered the candidates‟ age, gender, and overall attitudes in their ratings. All raters felt that the candidates were nervous during testing but only one of them said that he sympathized for it in his ratings. It is very interesting to find out that majority of the raters (four) declared that they compared a candidate with other candidates. Only one did not. One rater thought that English native speakers and ICAO level 6 were equivalent and vice versa. All raters said that they knew about the ICAO-required „cut-off‟ score and two of them accepted they considered the consequences of the candidates as pass/fail in their ratings. One rater considered changing the scores he already gave. Two raters considered himself as lenient while only one considered himself as a harsh rater.

The summary of the operational/trained raters‟ answers to the questionnaire is shown in Table 4.17 below:

Table 4.17: Operational/Trained Raters‟ answers to the questionnaire

	OT1	OT2	OT3	OT4	OT5
1. Gender	Male	Male	Male	Male	Male
2. Age (Years)	31-40	31-40	31-40	31-40	31-40
3. Educational level	B.A. (Education)	B.B.A. (Aerospace administration)	B.B.A. (General management)	B.Arch. (Architecture)	B.B.A. (Management)
4. Occupation	Pilot (Airline pilot)	Pilot (Airline pilot)	Pilot (Airline pilot)	Pilot (Airline pilot)	Pilot (Airline pilot)
5. Years of being in the occupation	1-5	11-15	6-10	11-15	6-10

6. First language(L1)	English	English	English	Thai	Thai
7. Duration of English study (years)	N.A.	N.A.	N.A.	11-15	>16
8. Level of English proficiency	N.A.	N.A.	N.A.	Good	Very good
9. Formal rater training and the course name(s)	Yes (3)	Yes (1)	Yes (3)	Yes (2)	Yes (3)
10. Exposure to various English native speakers" accents	Very much	Very much	Very much	Much	Very much
11. Exposure to Asian English accents	Much	Much	Very much	Very much	Very much
12. Exposure to European English accents	Some	Much	Very much	Some	Much
13. Degree of familiarity with linguistic terms	Little	Some	Much	Some	Some
14. Degree of familiarity with aviation operations and aeronautical communication	Much	Very much	Very much	Very much	Very much
15. Experience in language assessment	Some	Some	Very much	Some	Very much
16. Familiarity with using language descriptors	Some	Some	Much	Some	Much
17. Familiarity with ICAO language proficiency rating scale	Some	Some	Much	Much	Very much
18. Frequency of consulting the details of each ICAO descriptor in Doc. 9835 <u>before</u> listening to the speech samples	Frequently	Frequently	Always	Rarely	Frequently
19. Frequency of consulting the	Always	Frequently	Frequently	Rarely	Frequently

details of each ICAO descriptor in Doc. 9835 <u>during</u> listening to the speech samples					
20. Frequency of consulting the details of each ICAO descriptor in Doc. 9835 <u>after</u> listening to the speech samples	Always	Frequently	Always	Frequently	Frequently
21. Frequency of listening to the given speech samples <u>before</u> giving the final score	Always	Always	Sometimes	Sometimes	Sometimes
22. Frequency of taking notes <u>while</u> rating	Always	Always	Always	Frequently	Sometimes
23. Frequency of stopping the tapes for any reason <u>while</u> rating	Always	Frequently	Sometimes	Rarely	Sometimes
24. Frequency of stopping to listen for certain parts from the speech samples	Rarely	Frequently	Sometimes	Sometimes	Sometimes
25. Frequency of concentrating on errors made by the speaker	Sometimes	Sometimes	Frequently	Frequently	Frequently
26. Frequency of considering the relatedness/relevance of the content as a factor in your rating	Never	Sometimes	Frequently	Sometimes	Frequently
27. Having been busy lately?	Yes	Yes	Yes	Yes	Yes
28. Feeling bored/ exhausted/ tired during rating?	No	Yes	Yes	Yes	No
29. Having any	No	No	Yes	No	No

short term ailments?					
30. Having any long term ailments?	No	No	No	No	No
31. Having a good sleep/rest last night?	Yes	Yes	No	Yes	Yes
32. Had enough sleep/rest?	Yes	Yes	No	Yes	Yes
33. Was the room too cold?	No	No	Yes	No	Yes
34. Was the room too warm?	No	No	No	No	No
35. Was the room too dark?	No	No	No	No	No
36. Was the room too lighted?	No	No	No	No	No
37. Was the room too noisy?	No	No	Yes	No	No
38. Listening to the given speech sample from the beginning to the end without stopping at least once before rating?	Yes	No	No	Yes	Yes
39. Weighting each criterion equally before giving the final score?	Yes	Yes	Yes	No	Yes
40. Considering the quality of the content the candidates give as a factor in rating?	No	No	Yes	No	Yes
41. The test tasks were easy?	No	Yes	No	No	Yes
42. The test tasks were difficult?	No	No	No	No	No
43. The speech samples were too short?	No	No	No	Yes	No
44. The speech samples were too long?	No	No	No	No	No
45. Rating three speech samples consecutively was too much?	Yes	Yes	Yes	Yes	No

46. The interviewers/ interlocutors tried to help/ accommodate the candidate during the test?	No	Yes	No	Yes	No
47. The interviewers/ interlocutors performed their jobs appropriately/ effectively as they should have?	Yes	Yes	Yes	Yes	Yes
48. The interviewers/ interlocutors attempted to simplify their speech to facilitate the candidates or to match the candidates‟ level of language?	No	Yes	Yes	Yes	No
49. Considering the candidates‟ age in rating?	No	No	No	No	No
50. Considering the candidates‟ gender in rating?	No	No	No	No	No
51. Considering the global/overall attitudes of the candidates?	No	No	No	No	No
52. The candidates were nervous during the test?	Yes	Yes	Yes	Yes	Yes
53. Sympathize for that nervousness in rating?	No	No	No	Yes	No
54. Comparing the candidate with other candidates in rating?	No	Yes	Yes	Yes	Yes

55. Every English native speaker must also be at ICAO Level 6?	Yes	No	No	No	No
56. Being at ICAO Level 6 equivalent to being an English native speaker?	Yes	No	No	No	No
57. Knowing that the „cut-off" score for this ICAO assessment is level 4?	Yes	Yes	Yes	Yes	Yes
58. Considering the consequence of the candidates as pass or fail in rating?	No	No	Yes	Yes	No
59. Considering changing the scores already given to them?	No	No	No	Yes	No
60. Considering as being a lenient rater	No	Yes	No	No	Yes
61. Considering as being a harsh rater	No	No	No	Yes	No

(1) TOEIC language proficiency interviewer/rater training course in 2006

(2) TRAINAIR Standardized Training Packages (STPs) interviewer/rater training course organized by Civil Aviation Training Center (CATC) in 2009.

(3) TOEIC language proficiency interviewing/rating training course in 2006 & TRAINAIR Standardized Training Packages (STPs) interviewer/rater training course organized by Civil Aviation Training Center (CATC) in 2009.

The data taken from the operational/untrained rater questionnaires showed that, similar to the operational/trained raters, all operational/untrained raters were male. Three of them aged between 31 to 40 years while the other two were between 41-50 years. Four held master"s degrees. Among these, three had M.B.A. and the other had a Master of Landscape Architecture. The only operational/untrained rater graduated with a bachelor"s degree in veterinary medicine. Again, none was in the field of language or language-related. This

group of raters had a mixed experience in the occupation. The veteran one had 21 to 25 years of experience while one had 16 to 20 years, another had 6 to 10 years and the other two had just 1 to 5 years. All of their first languages were Thai. Three raters studied English for more than 16 years and the other two did it between 11 to 15 years. Three of them considered their English proficiency as „very good". The other two considered theirs as „good". None had any formal rater training.

Their levels of exposure to various English native speakers" accents were quite high as „much" (4) and only one considered his as „some". Their degrees of exposure to Asian English accents seemed to be higher as one „very much" and four „much". They appeared to be less familiar with the European English accents since two of them rated their degrees as „much" while the other two as „little". Only one thought of his as „some". Possibly because of the same reason as the operational/trained rater that they had no educational background in language or linguistic, two raters considered their degrees of familiarity with linguistic terms as „none", the other two as „little" and only one as „some". On the contrary, two raters in this batch referred to their familiarity with aviation operations and aeronautical communication as „very much", two as „much" and one as „some". When being asked about their experience in language assessment, two raters answered „little", one answered „some" and one „none". It was worth noting that an operational/untrained rater judged his experience in language assessment as „much". Their degrees of familiarity with using language descriptors were: three „little", one „some" and one „none". Three raters in this group put their familiarity with the ICAO language proficiency rating scale as „some", one as „little and one as „much".

Three operational/untrained raters answered that they „sometimes" consulted the details of each ICAO descriptor before listening to the speech samples. One did it „frequently" and another „rarely" did it. Two raters „frequently" checked the details during listening. The other two did it „sometimes" and another „rarely" did it. After listening three raters „frequently" turned to the details of ICAO descriptors. One did it „sometimes" and another „never" did it. Two raters „frequently" took notes while rating. One did it „sometimes"; one „always" did it and one „never" took notes at all. Two raters stopped the tape for a reason while rating. The other two „rarely" did it and there was a rater who „never"

stopped it. Three raters in this group „rarely" stopped to listen for certain parts of the samples. One did it „sometimes" while another „never" stopped it. Two raters accepted that they „frequently" concentrated on the errors made by the candidates. The other two did it „sometimes" and one „always" did it. The majority of three raters answered that they „always" considered the relevance of the content as a factor in their ratings while the other two did it „sometimes".

Three operational/untrained raters said that they had been busy lately while the other two said „no". Only one rater admitted that he felt bored/exhausted/tired during rating. The others said „no". None of the raters in this group had either short or long term ailments. All of them had good and enough rest before rating. No one complained about the setting as being too warm, too dark, or too lighted. Only one rater said that the room was noisy. Three raters did not listen to the speech samples from the beginning to the end without stopping at least once while the other two did. Every rater admitted that he weighted each criterion equally before giving the final score. All of them also accepted that they considered the quality of the content the candidates gave as a factor in their ratings. None of them thought that the test tasks were easy. However, one in five raters thought that the test tasks were difficult while the others did not think so. No one thought that the speech samples were too short. On the contrary, two thought that they were too long. Three raters felt that rating three speech samples consecutively was too much while the other two did not. All raters thought that the interlocutors performed their jobs appropriately, though two raters thought that they tried to help/accommodate the candidates and three raters thought that they tried to simplify the speech to facilitate the candidates.

One operational/untrained rater admitted that he considered the candidates" age in his rating but no one said that he considered the candidates" gender. Two raters accepted to consider the candidates" overall attitudes while the other three did not. Almost all (four) raters felt that the candidates were nervous during testing. Two of them said that they also sympathized for this in their ratings. The other two did not. The majority of the raters (four) declared that they compared one candidate with the other candidates. Only one did not. No rater thought that English native speakers and ICAO level 6 were equivalent and vice versa.

Four raters said that they knew about the ICAO-required „cut-off" score and the same raters accepted they considered the consequences of the candidates as pass/fail in their ratings. Only one rater neither knew nor considered it. Nobody considered changing the scores they already gave. Only one rater considered himself as lenient while three raters considered themselves as harsh raters.

The summary of the operational/untrained raters" answers to the questionnaire is shown in Table 4.18 below:

Table 4.18: Operational/Untrained Raters" answers to the questionnaire

	OU1	OU2	OU3	OU4	OU5
1. Gender	Male	Male	Male	Male	Male
2. Age (Years)	31-40	41-50	31-40	31-40	41-50
3. Educational level	M.B.A. (General admin.)	B.Vet.med. (Veterinary medicine)	M.B.A. (General management)	M.L.A. (Landscape architecture)	M.A.M. (Aviation management)
4. Occupation	Pilot (Airline pilot)	Pilot (Airline pilot)	Pilot (Airline pilot)	Pilot (Airline pilot)	Pilot (Airline pilot)
5. Years of being in the occupation	6-10	16-20	1-5	1-5	21-25
6. First language(L1)	Thai	Thai	Thai	Thai	Thai
7. Duration of English study (years)	11-15	>16	11-15	>16	>16
8. Level of English proficiency	Very good	Good	Very good	Very good	Good
9. Formal rater training and the course name(s)	No	No	No	No	No
10. Exposure to various English native speakers" accents	Some	Much	Much	Much	Much
11. Exposure to Asian English accents	Very much	Much	Much	Much	Much
12. Exposure to European English accents	Much	Much	Little	Little	Some
13. Degree of familiarity with linguistic	None	None	Some	Little	Little

terms					
14. Degree of familiarity with aviation operations and aeronautical communication	Some	Very much	Much	Very much	Much
15. Experience in language assessment	Little	Some	None	Much	Little
16. Familiarity with using language descriptors	Little	None	Little	Some	Little
17. Familiarity with ICAO language proficiency rating scale	Some	Some	Little	Much	Some
18. Frequency of consulting the details of each ICAO descriptor in Doc. 9835 <u>before</u> listening to the speech samples	Sometimes	Rarely	Sometimes	Sometimes	Frequently
19. Frequency of consulting the details of each ICAO descriptor in Doc. 9835 <u>during</u> listening to the speech samples	Sometimes	Sometimes	Frequently	Rarely	Frequently
20. Frequency of consulting the details of each ICAO descriptor in Doc. 9835 <u>after</u> listening to the speech samples	Sometimes	Frequently	Frequently	Never	Frequently
21. Frequency of listening to the given speech samples <u>before</u> giving the final score	Sometimes	Never	Never	Some-times	Frequently
22. Frequency of taking notes <u>while</u> rating	Frequently	Never	Frequently	Always	Sometimes

23. Frequency of stopping the tapes for any reason <u>while</u> rating	Sometimes	Never	Rarely	Rarely	Sometimes
24. Frequency of stopping to listen for certain parts from the speech samples	Sometimes	Never	Rarely	Rarely	Rarely
25. Frequency of concentrating on errors made by the speaker	Frequently	Always	Sometimes	Frequently	Sometimes
26. Frequency of considering the relatedness/relevance of the content as a factor in your rating	Sometimes	Always	Always	Always	Sometimes
27. Having been busy lately?	No	No	Yes	Yes	Yes
28. Feeling bored/ exhausted/ tired during rating?	No	Yes	No	No	No
29. Having any short term ailments?	No	No	No	No	No
30. Having any long term ailments?	No	No	No	No	No
31. Having a good sleep/rest last night?	Yes	Yes	Yes	Yes	Yes
32. Had enough sleep/rest?	Yes	Yes	Yes	Yes	Yes
33. Was the room too cold?	Yes	No	No	No	No
34. Was the room too warm?	No	No	No	No	No
35. Was the room too dark?	No	No	No	No	No
36. Was the room too lighted?	No	No	No	No	No
37. Was the room too noisy?	No	No	No	Yes	No
38. Listening to the given speech sample from the beginning to	No	Yes	No	Yes	No

the end without stopping at least once before rating?					
39. Weighting each criterion equally before giving the final score?	Yes	Yes	Yes	Yes	Yes
40. Considering the quality of the content the candidates give as a factor in rating?	Yes	Yes	Yes	Yes	Yes
41. The test tasks were easy?	No	No	No	No	No
42. The test tasks were difficult?	No	No	No	No	Yes
43. The speech samples were too short?	No	No	No	No	No
44. The speech samples were too long?	Yes	No	Yes	No	No
45. Rating three speech samples consecutively was too much?	No	Yes	No	Yes	No
46. The interviewers/ interlocutors tried to help/ accommodate the candidate during the test?	No	No	No	Yes	Yes
47. The interviewers/ interlocutors performed their jobs appropriately/ effectively as they should have?	Yes	Yes	Yes	Yes	Yes
48. The interviewers/ interlocutors attempted to simplify their speech to facilitate the candidates or	No	Yes	No	Yes	Yes

to match the candidates‟ level of language?					
49. Considering the candidates‟ age in rating?	No	No	No	Yes	No
50. Considering the candidates‟ gender in rating?	No	No	No	No	No
51. Considering the global/overall attitudes of the candidates?	No	No	Yes	Yes	No
52. The candidates were nervous during the test?	No	Yes	Yes	Yes	Yes
53. Sympathize for that nervousness in rating?	N.A.	No	No	Yes	Yes
54. Comparing the candidate with other candidates in rating?	Yes	No	Yes	Yes	Yes
55. Every English native speaker must also be at ICAO Level 6?	No	No	No	No	No
56. Being at ICAO Level 6 equivalent to being an English native speaker?	No	No	No	No	No
57. Knowing that the „cut-off‟ score for this ICAO assessment is level 4?	Yes	Yes	No	Yes	Yes
58. Considering the consequence of the candidates as pass or fail in rating?	Yes	Yes	No	Yes	Yes
59. Considering changing the scores already	No	No	No	No	No

given them?					
60. Considering as being a lenient rater	No	No	No	Yes	No
61. Considering as being a harsh rater	No	Yes	Yes	Yes	No

As for the raters‟ remarks or the statements of raters‟ opinions concerning the candidates‟ performance given by the raters, they were written by each rater after listening to each speech sample. They were grouped by each criterion and shown in Appendix F, G, and H.

Section Three: Results and discussion concerning the factors affecting the raters‟ decision-making

In this section, the content analysis was employed to examine the factors that might influence the raters‟ decision-making. Those factors are shown in the tables as follows:

Table 4.19 shows the educational backgrounds of the linguistic/trained raters if they were English or language related.

Table 4.19: Educational backgrounds of the linguistic/trained raters (LT)

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
1.English/language related	x		x		x		x		x		*LT1: "Bachelor of Arts majored in English."* *LT2: "I graduated with Bachelor of Arts majored in Linguistics and Master of Arts majoring in Teaching. Now I'm a Ph.D. candidate in Linguistics."* *LT3: "I graduated with a bachelor"s degree which is Bachelor of Arts majored in English and two master"s degrees – one is Master of Science majored in Education and the other is Master of Arts majored in English Teaching."* *LT4: "Bachelor of Education majored in English teaching."* *LT5: "I graduated with a bachelor"s degree in Education and a master"s degree in Linguistics. I am also currently a Ph.D. student in Education, majoring in higher education."*

All of the linguistic/trained raters graduated with at least a degree in English or linguistics or English-related i.e. English teaching. That means all of them have foundation of language and/or English. Two of them (LT1 and LT4) hold a bachelor"s degree while one has two master"s degrees (LT3) and the other two (LT2 and LT5) are studying for Ph.D.

Table 4.20 shows the educational backgrounds of the linguistic/untrained raters if they were English or language related.

Table 4.20: Educational backgrounds of the linguistic/untrained raters (LU)

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
1.English/language related	x		x		x		x		x		*LU1: "I graduated with Bachelor of Arts majoring in English. Master of Arts majored in Teaching English as a Second language. Now I'm a Ph.D. candidate in the English as an international language program majoring in Language Assessment and Evaluation."* *LU2: "Bachelor of Arts majored in English and Master of Arts majoring in Language and Culture for Communication. I also got a Diploma in Applied Linguistics. Now I'm a Ph.D. candidate in the English as an international language program majoring in English language instructions."* *LU3: "I graduated with a bachelor"s and a master"s degree in English language. Now I'm doing my Ph.D. in English as an international language program majoring in instructions."* *LU4: "I graduated with a bachelor"s degree in English teaching and a master"s degree in language assessment and evaluation. I'm now studying for a Ph.D. in the English as an international language program majoring in the same field - language assessment and evaluation."*

	LU5: "I am a Ph.D. candidate in the English as an international language program majoring in English teaching instructions and curriculum development. Before this I graduated with Master of Arts in Teaching English as a foreign language and a bachelor"s degree in Business English."

All of the linguistic/untrained raters are Ph.D. candidates in the English as an international language program. Three of them (LU2, LU3 and LU5) are majoring in teaching instructions while the other two (LU1 and LU4) are in language assessment and evaluation.

Table 4.21 shows the educational backgrounds of the operational/trained raters if they were English or language related.

Table 4.21: Educational backgrounds of the operational/trained raters (OT)

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
1.English/language related		x		x		x		x		x	*OT1: "Bachelor"s degree in Education majored in General education."* *OT2: "I graduated with Bachelor of Science majored in Aerospace Administration, specialized in Airline management and Airport management."* *OT3: "B.B.A. in General management"* *OT4: "I graduated with a bachelor"s degree in*

	Architecture. Before studying in the university, I was an AFS scholar studying in the USA for one year. It was in the state of Georgia." *OT5: "It was in management." "Yes. BBA"*

The raters in the operational/trained group graduated in three different fields. Three raters have knowledge in management. Two have BBA (Bachelor of Business Administration). Even though one rater has a bachelor"s degree of science, his major is also in management (Aerospace Administration). One has his background in architecture while another in education. It is worth noting that three raters in this category (OT1, OT2 & OT3) consider English as their first language and requested to conduct the interview in English. Another rater (OT5) did the same by requesting to speak English during the interview though his first language is Thai.

Table 4.22 shows the educational backgrounds of the operational/untrained raters if they were English or language related.

Table 4.22: Educational backgrounds of the operational/untrained raters (OU)

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
1.English/language related		x		x		x		x		x	***OU1:*** *"I graduated with Bachelor of Engineering majored in Electrical Engineering and also M.B.A. in General Management."* ***OU2:*** *"I graduated with a bachelor"s degree in Veterinary Medicine."* ***OU3:*** *"I graduated with a bachelor"s degree in Economics and a Master of Business Administration in General management from the States."*

	OU4: *"I graduated with both the bachelor"s and the master"s degrees in Landscape Architecture from Australia."* ***OU5:*** *"I graduated with bachelor"s degree in Chemistry from University of Liverpool in the U.K. and master"s degree in Aviation Management from Griffith University, Australia. It"s was distant learning."*

The batch of operational/untrained raters has background in various fields. Four raters (OU1, OU3, OU4 and OU5) graduated with master"s degrees. Even if three of them (OU1, OU3 and OU5) have their master"s in management i.e. two (OU1 and OU3) have MBA (Master of Business Administration) and one (OU5) specializes in aviation management, their first degrees are different. One (OU1) is in engineering, one (OU3) in economics while another (OU5) in chemistry. One (OU4) has both his master"s and bachelor"s degrees in architecture. The only rater with bachelor"s degree (OU2) graduated in veterinary medicine.

Table 4.23 shows the rating backgrounds of linguistic/trained raters if their education were rating related.

Table 4.23: Rating backgrounds of the linguistic/trained raters (LT)

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
2. Educational background is rating related		x	x		x		x		x		***LT1:*** *"No. It was just about English."* ***LT2:*** *"When I studied for my Master"s degree in Teaching, it had something to do with assessment." "It*

	was a subject that covered everything in assessment starting from test construction." "It was to assess students." ***LT3:*** *"Yes, to assess the students. It's something like what I'm doing now. Sometimes they were recorded and I rated them later."* ***LT4:*** *"Partly because we"d have some background in English language foundation so we can see if they have rigid basic English foundation. We studied sort of achievement test for students, not proficiency test like this."* ***LT5:*** *"It was sort of achievement test when I studied in education. I did that with my students but there was nothing concerned rating when I studied linguistics."*

None of the linguistic/trained raters has direct relationship of proficiency rating with their educational background. Though those who have background in teaching did a kind of assessment with their students, it was an achievement test – not proficiency test. They were trained to be raters in proficiency tests after their graduation.

Table 4.24 shows the rating backgrounds of the linguistic/untrained raters if their education were rating related.

Table 4.24: Rating backgrounds of the linguistic/untrained raters (LU)

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
2. Educational background is rating related	x		x		x		x		x		***LU1:*** *"Partly. I studied about the principles of rating."* ***LU2:*** *"Yes. Most of it concerned student assessment, a kind of achievement test."* ***LU3:*** *"Yes, in a way, for example those terms used in the descriptors such as ,,style". If I didn"t study in linguistics, I wouldn"t know what ,,style" means or what the correct grammar is." "But I've never done rating like those studying in teaching who used to assess their students."* ***LU4:*** *"Yes, in terms of assessment."* ***LU5:*** *"No, not directly." "In the track of teaching, I studied mostly in teaching curriculum and teaching methodology. I studied some in assessment but not in details, just to know, okay, this is holistic, what the scale is, but not the test design or sort of things like that." "We tried to make most of our practical student assessment in the form of objective assessment because it affected the student grades." "If it is subjective assessment, it would be a formal test because it would be difficult to assess."*

Despite the fact that two of the linguistic/untrained raters (LU1 and LU4) are Ph.D. students in language assessment and evaluation, they never conducted any kind of proficiency rating before. They just have some knowledge in the assessment principles. For the other three who are Ph.D. students in the field of instructions (LU2, LU3 and LU5), one (LU2) did some kind of achievement

testing while another (LU5) *"studied some in assessment but not in details*". Another rater (LU3) just understands some terms used in the descriptors e.g. „style". The rating they did in this study was their first „real time" rating.

Table 4.25 shows the rating backgrounds of the linguistic/trained raters if their education were rating related.

Table 4.25: Rating backgrounds of the operational/trained raters (OT)

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
2. Educational background is rating related		x		x		x		x		x	***OT1:*** *"No, not rating."* ***"****I had to develop a test but it was mainly for elementary school level. It was quite relaxed. We had to take a course in assigning the proper value according to elementary school or high school or grade level they were trying to specialize in or to be in that." "I felt more comfortable when it was the absolute numbers or the one done by percentage whether right or wrong." "I felt uncomfortable when it came to more of the opinion-based assessment."* ***OT2: "****No, not at all."* ***OT3:*** *"I don"t think so."* ***OT4:*** *"No."* ***OT5:*** *"No. It was definitely different scope, different field." "At least we had some assessments of the students but not about the language. It was some other kind of assessment."*

In spite of being the operational/trained raters, none of the raters in this family has relationship between their educational background and the rating. All of them were trained to be raters afterwards.

Table 4.26 shows the rating backgrounds of the operational/untrained raters if their education were rating related.

Table 4.26: Rating backgrounds of the operational/untrained raters (OU)

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
2. Educational background is rating related		x		x		x		x		x	***OU1:*** *"No."* ***OU2:*** *"No."* ***OU3:*** *"No."* ***OU4:*** *"No. Nothing concerns rating."* ***OU5:*** *"No."*

Similar to the operational/trained raters, the operational/untrained raters have no relationship between their educational background and rating. The difference between them is that the operational/untrained raters never get any kind of rater training.

Table 4.27 shows the mental conditions of the linguistic/trained raters if they were busy lately.

Table 4.27: Mental conditions affected by being busy of the linguistic/trained raters (LT)

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
3. Being busy lately	x		x		x		x		x		***LT1:*** *"Yes. I was busy checking and scoring the ATT pilot English exam papers." "If it has any effect on this rating, I think it has a positive effect because it makes me more familiar with the rating process especially with the criteria such as structure and vocabulary. Even though they"re not exactly the same, they still have something in common. That makes my today rating faster."* ***LT2:*** *"Yes, very busy (laughter)." "Both at home and in the office."* ***LT3:*** *"There are many courses now. I teach around three days a week. It"s busy as usual."* ***LT4:*** *"Yes, as usual. Most of them are job-related such as budgeting."* ***LT5:*** *"Yes. It"s about my routine work. Today is the first day of my present course."*

All of the linguistic/trained raters admitted that they were busy with either or both their routine jobs and other personal business e.g. their families.

Table 4.28 shows the mental conditions of the linguistic/untrained raters if they were busy lately.

Table 4.28: Mental conditions affected by being busy of the linguistic/untrained raters (LU)

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
3. Being busy lately	x		x			x	x		x		***LU1:*** *"Yes. My two dogs that I love very much died in a row, one in April and the other in May. I wasn"tactually busy but it was rather tragic to me."* ***LU2:*** *"Yes. Both teaching and conducting research."* ***LU3:*** *"No, not at all. I'm just dealing with my study, my research."* ***LU4:*** *"Yes, with my study. There are a lot of reading and writing. I'm married but haven't got any child so now I'm spending most of my time with my Ph.D. study."* ***LU5:*** *"Very busy both my administrative duties and my Ph.D. study."*

Almost all of the raters (LU2, LU4 and LU5) in the group of linguistic/untrained said that they were busy with their study. One rater (LU1) stated her grief over the death of her two dogs. Only one rater (LU3) did not say that she was busy.

Table 4.29 shows the mental conditions of the operational/trained raters if they were busy lately.

Table 4.29: Mental conditions affected by being busy of the operational/trained raters (OT)

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
3. Being busy lately	x		x		x		x		x		***OT1:*** *"I've been busy, not at work but at family oriented, not work-related."* ***OT2:*** *"Sort of busy."* ***OT3:*** *"Yes." "Family and relatives. Family business."* ***OT4:*** *"Yes, quite busy. I'm under the process of pre-evaluation to be a captain candidate."* ***OT5:*** *"Busy? Yes. Well, I switch between flying and interviewing. So on the days off I'd come in and do the interviews. This month I fly almost a hundred hours. Right now it"sabout eighty-five and I have two more flights to go." "Just for this month because last month they took two weeks off of my schedule for attending the course."*

All operational/trained raters said that they were busy with something. If not their flight duties, it was their family business.

Table 4.30 shows the mental conditions of the operational/untrained raters if they were busy lately.

Table 4.30: Mental conditions affected by being busy of the operational/untrained raters (OU)

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
3. Being busy lately		x		x	x		x		x		***OU1:*** *"Not quite. Not much flying."* ***OU2:*** *"No. There"s nothing speaal."* ***OU3:*** *"Yes, with some personal business. I'm going to get married. I"mpreparing for my wedding ceremony."* ***OU4:*** *"Yes. I just moved to my new house recently. My wife is pregnant and I'm expecting a baby. My younger brother just got married and we had his wedding ceremony. Yes, I'm quite busy." (Laughter)* ***OU5:*** *"Busy but not specially. Just some personal business such as getting the car repaired and household activities like fixing the broken water pipe."*

Two of the raters (OU1 and OU2) in the operational/untrained family did not think that they were busy. The other three (OU3, OU4 and OU5) were busy with their family matters.

Table 4.31 shows the mental conditions of the operational/trained raters if they returned from their last flight more than 24 hours.

Table 4.31: Mental conditions affected by their last flights of the operational/trained raters (OT)

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
4. Returning from his last flight more than 24 hours	x			x	x			x		x	***OT1:*** *"Five days ago." "That was Munich."* ***OT2:*** *"It was yesterday at 6.30 p.m. from Kathmandu."* ***OT3:*** *"Last flight I went to Oslo and came back four days ago."* ***OT4:*** *"Just at 7 a.m. this morning."* ***OT5:*** *"Last night from Brisbane." "It arrived at 8.35 p.m."*

Two raters (OT1 and OT3) in the operational/trained group returned from their last flights more than three days ago. The other two (OT2 and OT5) came back the day before the rating. Only one rater (OT4) just returned from his last flight in the morning of the rating day.

Table 4.32 shows the mental conditions of the operational/untrained raters if they returned from their last flight more than 24 hours.

Table 4.32: Mental conditions affected by their last flights of the operational/untrained raters (OU)

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
4. Returning from his last flight more than 24 hours	x		x		x		x			x	***OU1:*** *"Yesterday."* ***OU2:*** *"Yesterday morning from Chiangmai."* ***OU3:*** *"Two days ago from Kansai Osaka."* ***OU4:*** *"It was more than a week ago, almost two weeks because I asked for a leave to help prepare my brother's wedding ceremony."* ***OU5:*** *"It was yesterday afternoon from Kuala Lumpur."*

Three operational/untrained raters (OU1, OU2 and OU5) returned from their last flights a day before the rating. One (OU3) arrived two days ago and the other one (OU4) came back *"almost two weeks"* ago.

Table 4.33 shows the mental conditions of the linguistic/trained raters if they felt bored/exhausted/tired during rating.

Table 4.33: Mental conditions affected by their boredom of the linguistic/trained raters (LT)

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
5. Feeling bored/exhausted/tired	x		x			x		x	x		***LT1:*** *"Yes but it wasn't the kind of physical tiredness. It was rather the kind of mental tiredness that I had to comply with ICAO criteria and*

during rating	*requirements which are clearly stated. I had to use a lot of energy for that which made me feel tired. Actually it was more like mental fatigue."* ***LT2:*** *"It was quite exhausted but not much because I'm quite familiar with listening for this rating purpose."* ***LT3:*** *"No, it was neither tiring nor boring because I'm used to this kind of job but my ears were hurt. I wore the headphone too long but it wasn"t a problem." "Maybe a little exhausted because I didn"tget up at all. Usually when I do the rating in my office, I can have some snacks and walk around and I normally rate two samples and then relax a while but it was three samples. However, I went to the toilet once."* ***LT4:*** *"No, not at all because I got involved with it."* ***LT5:*** *"What should I say? Let"s say I prefer doing something else. I'm not happy rating so it made me feel sick. Rating many people in a row also made me feel tired. It's more of boring than tiring because I wasn"t happy doing it."*

Two linguistic/trained raters (LT2 & LT3) said that they were familiar with this kind of rating. So one of them (LT2) *"was quite exhausted but not much"* while another (LT3) *"was neither tiring nor boring*". One rater (LT4) denied any tiredness "*because I got involved with it."* LT1 was not „physically" tired but she was „mentally" tired because she used a lot of her energy *"to comply with*

ICAO criteria and requirements which are clearly stated". LT5 was unique among the others to say that she "*prefers doing something else"* and *"It"s more of boring than tiring because I wasn"t happydoing it"*.

Table 4.34 shows the mental conditions of the linguistic/untrained raters if they felt bored/exhausted/tired during rating.

Table 4.34: Mental conditions affected by their boredom of the linguistic/untrained raters (LU)

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
5. Feeling bored/exhausted/tired during rating		x	x			x	x		x		***LU1:*** *"Not really because I never did this before so it was sort of fun." "It was quite interesting."* ***LU2:*** *"Yes because I wasn"tsure how long it would take and I wasn"tfamiliar with the test content. It was also because I didn"thave confidence in myself that I didn"t fully understand and then if I could rate them correctly. I felt quite uncomfortable doing it." "I felt tired not because it took so long but because I never saw the test before. The first time I listened to the samples was also the time I had to rate them." "Rating three candidates was not tiring."* ***LU3:*** *"No, I was neither bored nor tired but it"s like it was very long that some part they did good but dropped in the next part. After finishing I wasn"t sure if they did fine or not. So I had to go back and listened again." "But I didn"tfeel bored or tired. I could keep doing it." "I was sort of confused so I couldn"t make up my mind how I should have*

awarded the scores. Should it be „thræ" or „four"? Something like that." "Especially for the third candidate, he seemed to be fluent in the first part so I expected that he should have done that well in the interview part but he happened to be unable to convey his ideas smoothly. It looked like he had problems with his grammar so he uttered unevenly. That made me confused. Should I give him „thræ" or „four"?"

LU4: *"No, I didn"t feel bored rating these three candidates but I felt a little tired starting from the first candidate because there were a lot of technical terms. I'm unfamiliar with the test content. I have just an overview picture of the pilot"sjobs but when it comes to the point that gets deep in the details which involves many technical terms I couldn"t catch it. And I also wasn"tsure about the meanings. That made me tired. But I don"t think it"s hard for raters with more experience. This happened to be my first time so it was tough for me. Even though I study the manual and the rubrics before rating it was still tough."*

LU5: *"It was boring in the beginning because I didn"tknow how to assess. But after listening to the second guy and comparing with the first, I began to visualize more clearly how they differed." "I didn't know what they were about in the beginning because they were full of terms that I didn"tknow. I wasn"tsure if I could assess them because I didn"t know the vocabularies. But when I listened to the*

	second candidate, I knew that at least I could rate them in terms of language. I started to concentrate more when listening to the second and the third candidates. I started to know more and could compare them more." "I was confused with the first guy because there were lots of technical terms but it was better when I rated the second and the third." "It wasn"tboring with the first one but it was no fun because I didn"t understand." "I wasn"t tired listening but I was tired from my journey. If I could start listening right in the morning without the necessity of traveling, it would be better."

All linguistic/untrained raters had different perspectives in the ratings in that they felt differently from the same thing. LU1 did not really feel tired or bored because "*I never did this before so it was sort of fun."* while LU2 said „yes" because "*I wasn"tsure how long it would take and I wasn"tfamiliar with the test content.*" Four raters complained about their unfamiliarity with the test content and the technical terms used in the test because of their lack of background in aviation. It may be concluded that it was neither tiredness nor boredom that directly affected their ratings but it might be rather their unfamiliarity with the test content and the technical terms used in the test because of their lack of background in aviation that affected their ratings.

Table 4.35 shows the mental conditions of the operational/trained raters if they felt bored/exhausted/tired during rating.

Table 4.35: Mental conditions affected by their boredom of the operational/trained raters (OT)

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
5. Feeling bored/exhausted/tired during rating		x	x		x		x			x	***OT1:*** *"Because we did not discuss what exactly we"regonna do today so I did not prepare myself." "I spent almost an hour reviewing the ICAO proficiency level scales to visualize them." "I'm confident that I got the right score for the sample but I'll be more comfortable having more time reviewing the material."* ***OT2:*** *"I'd say the first was okay. The first was the best one. It was interesting what was said and the second one was weaker than the first one. So I tried to catch what"swrong, what the speaker was trying to say, all the different criteria. So you know what the questions are but you try to grab that. It tends to get boring at the end. And… the first was interesting and I didn"tknow what"sgonna happen about the whole subject things. I was also like being involved, you know, the person being interviewed and also the rater and then he talked more often than the second and the third." "Sometimes I wasn"tjust listening to the interviewee, I was trying to catch the interviewer because on the third one, he tended to help the interviewee by rephrasing the questions. I don"tknow. It could be just me or …." "That"s what happens when you get bored, you'd find something else to do." "If it"s all the same format, maybe I have to change my way of*

listening." "If they"reall different, it could be more interesting for me."

OT3: *"Yes. I was bored because it"s a ... rating is not a fun job. That"s why I was bored. It didn"t mean I didn"tpay attention to but it wasn"ta fun so when it wasn"t a fun, it may be bored." "Tiring would be a better word." "It was tiring because certain jobs like my new work, you know ... nobody gets tired but this one is just constantly concentrating on listen to parts of speech. Try to catch each part then try to think what"s the appropriate ICAO rating. That"s sort of a stress so ... yeah ... that"s the reason for being tiring and ... not quite so much fun." (Laughed).*

OT4: *"Rating the first one was still okay. I started to feel bored when rating the second and the third." "It was a mixture of boring and tiring. It was more on boring than tiring." "I was just a little tired." "I started to get bored while listening to the second sample. It was like the first guy. So my concentration was decreased because I already knew what would go on after that. Same thing happened when I listened to the third guy."*

OT5: *"During rating? No because last night I went to sleep early since I came back and I felt quite tired so I slept early and I woke up around 11. Came here, have lunch and I was ready for the rating."*

Contrary to the linguistic/untrained raters, the operational/trained raters did not have any problem with the test content or the technical terms. None of them felt tired but some of them (OT2, OT3 and OT4) got rather bored of rating.

Table 4.36 shows the mental conditions of the operational/untrained raters if they felt bored/exhausted/tired during rating.

Table 4.36: Mental conditions affected by their boredom of the operational/untrained raters (OU)

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
5. Feeling bored/exhausted/tired during rating	x		x		x		x		x		***OU1:*** *"A little when starting to rate the third candidate because it became repetitive. It was sort of tiring." "There were long pauses after the questions." "It might be boring because it was recurring."* ***OU2:*** *"The first two were okay but it was sort of tiring when listening to the third guy because I had to think, to pay attention, to focus all the time. It was different from relax listening."* ***OU3:*** *"I got a little bored in the part of radiotelephony. The personal interview part was more interesting. I enjoyed listening to that part." "I think I was quite tired because I tried to do it as precise as I could. I concentrated a lot."* ***OU4:*** *"No, not bored but I felt exhausted. I gave less time to the third candidate than the first one. But I got all details. However, the third sample*

	happened to be shorter than the others." "I spent quite some time for the first two candidates, so I started to get exhausted listen to the third. I think I also had problems writing the reasons why I awarded the scores to them. Thinking what and how to write made me exhausted. If I don"thave to write, just rate them, I guarantee I can rate ten persons today. I mean just fill the scores in the tables." ***OU5:*** *"No, not bored but a little tired because I had to focus on the listening and categorizing each candidate."*

All of them said that they felt tired. Three out of five operational/untrained raters (OU1, OU2 & OU4) said that they got tired when rating the third candidate.

Table 4.37 shows the mental conditions of the linguistic/trained raters if they had any incident on the way to rating.

Table 4.37: Mental conditions affected by any incident of the linguistic/trained raters (LT)

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
6. Any incident on the way to rating		x		x		x		x		x	***LT1:*** *"No."* ***LT2:*** *"No. The traffic was good."* ***LT3:*** *"No, everything was normal."*

	LT4: *"No. I came here by the company bus as usual. Nothing exciting happened."* ***LT5:*** *"No."*

None of the raters in the linguistic/trained group experienced any kind of incident on their way to rating.

Table 4.38 shows the mental conditions of the linguistic/untrained raters if they had any incident on the way to rating.

Table 4.38: Mental conditions affected by any incident of the linguistic/untrained raters (LU)

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
6. Any incident on the way to rating		x		x		x	x		x		***LU1:*** *"No."* ***LU2:*** *"No."* ***LU3:*** *"No."* ***LU4:*** *"Yes. My car broke down and I had to take the subway to get to see you. It was quite a rush."* ***LU5:*** *"It was hot. Maybe because I worked in the morning and it was very busy. Then I came here."*

Three raters (LU1, LU2 and LU3) in this linguistic/untrained category were not bothered by any means on their way to rating. One rater (LU4) might be irritated by the incident of his car breakdown. The last rater (LU5) was not exasperated by any incident but he complained about the hot weather and his busy day which might be annoying enough to make him mention about them.

Table 4.39 shows the mental conditions of the operational/trained raters if they had any incident on the way to rating.

Table 4.39: Mental conditions affected by any incident of the operational/trained raters (OT)

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
6. Any incident on the way to rating		x		x	x			x		x	***OT1:*** *"No because I understand what these people were trying to do, how important the material was to these people so I wasn"t tired at all."* ***OT2:*** *"No."* ***OT3:*** *"Well ... maybe in the morning. The traffic was annoying. They stopped the traffic from the express way."* ***OT4:*** *"No."* ***OT5:*** *"None. Nothing unusual. Just mostly cloudy sky. Other than that, everything was normal."*

Almost all operational/.trained raters (OT1, OT2, OT4 and OT5) were not annoyed by anything on their way to rating. Only one (OT3) complained about the traffic which was *"annoying"*.

Table 4.40 shows the mental conditions of the operational/untrained raters if they had any incident on the way to rating.

Table 4.40: Mental conditions affected by any incident of the operational/untrained raters (OU)

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
6. Any incident on the way to rating		x		x		x		x		x	***OU1:*** *"No. I live near here."* ***OU2:*** *"I usually drive fast but there was nothing special."* ***OU3:*** *"It was quite a routine. The traffic wasn"tso bad. It was a low stress today."* ***OU4:*** *"No. I was just worried that I might have come late. So I rushed to get here. I forgot about the parking space problem. You know it"s hard to find one in our headquarters."* ***OU5:*** *"No."*

None of the operational/untrained rater was annoyed by any means on their way to rating. However, one rater (OU4) said that it was rather hard for him to find a parking space. He seemed to have some problem to find it. This might somewhat annoy him.

Table 4.41 shows the mental conditions of the linguistic/trained raters if they had any short-term aliment.

Table 4.41: Physical conditions in terms of short-term ailments of the linguistic/trained raters (LT)

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
7. Short-term ailments		x		x	x			x		x	***LT1:*** *"No. I felt 100% ready to do the rating."* ***LT2:*** *"No."* ***LT3:*** *"No, but I'm normally allergic when being in an air-conditioned room, allergic to dust from the air-conditioner."* ***LT4:*** *"No."* ***LT5:*** *"No."*

Almost all linguistic/trained raters did not have any short-term ailments during their ratings. However, LT3 accepted that she was usually allergic to dust from the air-conditioner. This might have some effect on her rating since it was conducted in an air-conditioned room.

Table 4.42 shows the mental conditions of the linguistic/untrained raters if they had any short-term aliment.

Table 4.42: Physical conditions in terms of short-term ailments of the linguistic/untrained raters (LU)

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
7. Short-term ailments		x		x		x		x		x	***LU1:*** *"No."* ***LU2:*** *"No."* ***LU3:*** *"No."* ***LU4:*** *"No."* ***LU5:*** *"No."*

All raters in the linguistic/untrained did not have any short-term ailments during their ratings. Thus, this factor should have not affected their ratings.

Table 4.43 shows the mental conditions of the operational/trained raters if they had any short-term aliment.

Table 4.43: Physical conditions in terms of short-term ailments of the operational/trained raters (OT)

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
7. Short-term		x		x	x			x		x	***OT1:*** *"No."*

ailments	***OT2:*** *"No."* ***OT3:*** *"I got a headache already, not from this but from the seat was not comfortable. I had to lean too far back. I tried to find the knob that makes it straight up. I couldn"t find it so I'm on the leaning like this. Yeah … I got a back pain and headache." "I had a bit of cold two days ago. So … just about to finish, hopefully." "The only way it would affect the rating may be the thought I'm in a bit pain, not pain but discomfort from uncomfortable seat." "It makes you think „Try to keep this up and finish it so no more back pain"." "It's the bad seat." "It's not from the rating." "No. I don"t think it affects my rating in any way."* ***OT4:*** *"No."* ***OT5:*** *"During rating? No. If I sat longer, maybe."*

Four operational/trained raters (OT1, OT2, OT4 andOT5) said that they did not have any short-term aliments during their ratings. OT3 was the only rater in this group who complained quite a lot about "*the seat was not comfortable"* and it made him "*got a back pain and headache*". However, he finally confirmed that he did not think it affected his rating in any way. Therefore, this factor should not have effect on their ratings.

Table 4.44 shows the mental conditions of the operational/untrained raters if they had any short-term aliment.

Table 4.44: Physical conditions in terms of short-term ailments of the operational/untrained raters (OU)

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
7. Short-term ailments		x		x		x	x			x	***OU1:*** *"No."* ***OU2:*** *"No."* ***OU3:*** *"No. I'm in good shape. I just went to a fitness center yesterday."* ***OU4:*** *"I got a little pain on my left ankle because I played soccer with my friends yesterday."* ***OU5:*** *"No."*

Four out of five operational/untrained raters (OU1, OU2, OU3 and OU5) stated that they did not have any short-term ailments during their ratings. Just OU4 said that he *"got a little pain*" on his left ankle which might have some effect on his rating.

Table 4.45 shows the mental conditions of the linguistic/trained raters if they had a good sleep/rest the night before rating.

Table 4.45: Physical conditions in terms of a good sleep/rest of the linguistic/trained raters (LT)

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
8. Having a good sleep/rest the night before rating	x		x		x		x			x	***LT1:*** *"Yes."* ***LT2:*** *"Yes, around five hours."* ***LT3:*** *"Yes, I went to sleep from 10 p.m. until 6 a.m."* ***LT4:*** *"Yes. I usually sleep just four or four hours and a half. I routinely go to bed at midnight and get up around 4 or 4.30 in the morning."* ***LT5:*** *"No. I slept only three hours last night."*

Almost all linguistic/trained raters (LT1, LT2, LT3 and LT4) mentioned that they slept well the night before rating, except LT5 who claimed that she slept *"only three hours"*. Having not a good sleep might affect her rating.

Table 4.46 shows the mental conditions of the linguistic/untrained raters if they had a good sleep/rest the night before rating.

Table 4.46: Physical conditions in terms of a good sleep/rest of the linguistic/untrained raters (LU)

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
8. Having a good sleep/rest the night before rating	x		x		x		x		x		***LU1:*** *"Yes."* ***LU2:*** *"Yes. It was around six hours."*

	LU3: *"Yes, I slept well."* ***LU4:*** *"Yes. I slept from eleven until six in the morning."* ***LU5:*** *"I slept for seven hours. It was okay." "Usually I go to bed around eleven p.m. or midnight and get up around seven in the morning."*

All linguistic/untrained raters expressed that they had good sleep the night before rating. Hence, they should not have been affected by this factor in their ratings.

Table 4.47 shows the mental conditions of the operational/trained raters if they had a good sleep/rest the night before rating.

Table 4.47: Physical conditions in terms of a good sleep/rest of the operational/trained raters (OT)

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
8. Having a good sleep/rest the night before rating		x	x			x	x		x		***OT1:*** *"No, not at all."* ***OT2:*** *"Yes."* ***OT3:*** *"No because of the kids." They slept on my bed."* ***OT4:*** *"I managed to get some rest during the flight. After arriving home from the airport, I slept for two hours." "Let me say it was seven out of ten score."* ***OT5:*** *"Yeah. Quite a very good sleep." "I slept for eleven hours."*

OT1 and OT3 clearly stated that they did not sleep well the night before rating. This might suggest that they were affected by this sub-themes on their ratings. On the contrary, OT2 and OT5 said that they had good sleep. OT4 who just returned from his last flight in the morning of the rating told that he *"managed to get some rest during the flight"* He also "*slept for two hours after arriving home from the airport"*. He mentioned his sleep as *"seven out of ten score"* which could be concluded as „not enough" and it might affect his rating.

Table 4.48 shows the mental conditions of the operational/untrained raters if they had a good sleep/rest the night before rating.

Table 4.48: Physical conditions in terms of a good sleep/rest of the operational/untrained raters (OU)

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
8. Having a good sleep/rest the night before rating	x		x		x		x		x		***OU1:*** *"Yes. I got eight hours."* ***OU2:*** *"Yes. I slept from eleven until seven thirty in the morning."* ***OU3:*** *"Yes because I knew that I had to do this job today."* ***OU4:*** *"It was normal, around seven hours starting from 11 p.m. until 6 a.m. That"s my normal sleep pattern."* ***OU5:*** *"Yes. I usually sleep around nine hours each night."*

All operational/untrained raters committed that they had good sleep the night before rating. This could mean that it did not affect their ratings.

Table 4.49 shows the mental conditions of the linguistic/trained raters if they had enough rest/sleep the night before rating.

Table 4.49: Physical conditions in terms of an adequate sleep/rest of the linguistic/trained raters (LT)

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
9. Having enough rest/sleep the night before rating	x		x		x		x			x	***LT1:*** *"Yes."* ***LT2:*** *"Yes. Usually I sleep around that. I go to bed around 10 p.m. and get up around 5 or 6 a.m. so it was my standard sleeping."* ***LT3:*** *"Yes."* ***LT4:*** *"Yes."* ***LT5:*** *"No. Normally I sleep around six hours but I'm studying for Ph.D. That"swhy I go to bed around 2 or 3 a.m. and get up around 6 in the morning because I have to study every night before going to bed."*

LT5 was the only linguistic/trained rater who protested that she did not have adequate sleep the night before her rating and it might affect her rating. The other four raters admitted that they had enough of it.

Table 4.50 shows the mental conditions of the linguistic/untrained raters if they had enough rest/sleep the night before rating.

Table 4.50: Physical conditions in terms of an adequate sleep/rest of the linguistic/untrained raters (LU)

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
9. Having enough rest/sleep the night before rating	x		x		x		x		x		***LU1:*** *"Yes. I had quite a deep sleep."* ***LU2:*** *"Yes. I usually sleep around six hours."* ***LU3:*** *"Yes. I usually go to bed around 11 p.m., but it was midnight last night, and get up around 8 in the morning."* ***LU4:*** *"Yes. I usually sleep around six to eight hours each night. Last night I got seven hours so it was alright."* ***LU5:*** *"Yes."*

All five linguistic/untrained raters said that they had enough sleep the night before their ratings. Consequently, this should not have affected their ratings.

Table 4.51 shows the mental conditions of the operational/trained raters if they had enough rest/sleep the night before rating.

Table 4.51: Physical conditions in terms of an adequate sleep/rest of the operational/trained raters (OT)

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
9. Having enough rest/sleep the night before rating	x		x		x			x	x		***OT1:*** *"Really good sleep."* ***OT2:*** *"I think I did." "I usually go to bed around 10 p.m. and get up around 6 or 6.30 in the morning." "Last night I went to bed at 8.30. So it"s morethan enough."* ***OT3:*** *"Normally I sleep about six hours. Last night I slept for … about six but it wasn"ta good sleep."* ***OT4:*** *"Normally I sleep for three hours after each flight so it was a little less than usual." "I usually sleep around five to eight hours when I"mat home." "No, it was not enough to do work. But it had very little effect on my rating."* ***OT5:*** *"Normally I have kind of sleeping problem or sleeping disorder where I would wake up every two to three hours. Just wake up. Come and sit or sit up and then just go back to sleep again. It"s a discontinued sleep. That affects on some days. Not last night. Last night I had a full twelve hour sleep."*

Almost all operational/trained raters but OT4 accepted that they had adequate sleep the night before their ratings. Even OT5, who explained that he usually has *"kind of sleeping problem or sleeping disorder",* said that he had an unusual *"full twelve hour sleep".* As a result, these four raters should not have been affected by this sub-themes on their ratings. Nonetheless, OT4 who just

returned from his last flight in the morning of the rating admitted that his sleep "*was not enough to do work",* though "*it had very little effect on my rating".* This could still be inferred that he was somewhat affected by having „not enough" sleep on his rating.

Table 4.52 shows the mental conditions of the operational/untrained raters if they had enough rest/sleep the night before rating.

Table 4.52: Physical conditions in terms of an adequate sleep/rest of the operational/untrained raters (OU)

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
9. Having enough rest/sleep the night before rating	x		x		x		x		x		***OU1:*** *"Yes. I slept from midnight to eight in the morning."* ***OU2:*** *"Yes."* ***OU3:*** *"I slept for eight hours from ten to six in the morning which was more than usual. Usually I sleep around six to seven hours each night." "Probably because I worked out for almost half day in the gym so I went to bed early."* ***OU4:*** *"Yes."* ***OU5:*** *"Yes. I went to bed at 8.30 p.m. until 5.30 a.m."*

None of the operational/untrained raters complained about having inadequate sleep the night before their ratings. Therefore, they should not have been affected by this factor on their ratings.

Table 4.53 shows the physical settings if the linguistic/trained raters felt the room was too warm or too cold or neither.

Table 4.53: Physical settings in terms of the room temperature felt by the linguistic/trained raters (LT)

Sub-themes	Raters															Meaning units
	LT1			LT2			LT3			LT4			LT5			
	W	C	N	W	C	N	W	C	N	W	C	N	W	C	N	
10. The room temperature was too warm (W) or too cold (C) or neither (N)			x			x			x			x			x	***LT1:*** *"No. But there were some mosquitoes underneath the desk. It was a little annoying."* ***LT2:*** *"No. It was fine."* ***LT3:*** *"No. It was neither too cold nor too warm."* ***LT4:*** *"No."* ***LT5:*** *"No."*

None of the linguistic/trained raters said that the room they conducted their ratings was too cold or too warm. It might be concluded that the room temperature did not affect their ratings. However, LT1 complained that she was annoyed by another factor which was not expected or included in the interview. They were *"some mosquitoes underneath the desk"*. This annoyance might have some effect on her rating.

Table 4.54 shows the physical settings if the linguistic/untrained raters felt the room was too warm or too cold or neither.

Table 4.54: Physical settings in terms of the room temperature felt by the linguistic/untrained raters (LU)

Sub-themes	Raters															Meaning units
	LU1			LU2			LU3			LU4			LU5			
	W	C	N	W	C	N	W	C	N	W	C	N	W	C	N	
10. The room temperature was too warm (W) or too cold (C) or neither (N)			x			x		x				x			x	***LU1:*** *"No. It was quite comfortable."* ***LU2:*** *"No. it was alright when I put the jacket on."* ***LU3:*** *"It was a bit cool occasionally."* ***LU4:*** *"Okay. It was fine."* ***LU5:*** *"It was alright, not too warm, not too cold."*

Almost all of the linguistic/untrained raters were annoyed by the room temperature. Only LU3 stated that *"it was a bit cool occasionally"*. This might *"occasionally"* affect her rating.

Table 4.55 shows the physical settings if the operational/trained raters felt the room was too warm or too cold or neither.

Table 4.55: Physical settings in terms of the room temperature felt by the operational/trained raters (OT)

Sub-themes	Raters															Meaning units
	OT1			OT2			OT3			OT4			OT5			
	W	C	N	W	C	N	W	C	N	W	C	N	W	C	N	
10. The room temperature was too warm (W) or too cold (C) or neither (N)			x			x		x				x		x		***OT1:*** *"No, it"s fine."* ***OT2:*** *"No."* ***OT3:*** *"Too cold."* ***OT4:*** *"No."* ***OT5:*** *"The room was quite cold, I guess because I think there were just two of us here in the office. If we had more people, then may be the room may be nicer." "It was a bit cold."*

Two operational/trained raters remarked that the room was cold. It was *"too cold"* for OT3 and *"quite cold"* for OT5. This might have an impact on their ratings. The others did not have any problem with the room temperature.

Table 4.56 shows the physical settings if the operational/untrained raters felt the room was too warm or too cold or neither.

Table 4.56: Physical settings in terms of the room temperature felt by the operational/untrained raters (OU)

Sub-themes	Raters															Meaning units
	OU1			OU2			OU3			OU4			OU5			
	W	C	N	W	C	N	W	C	N	W	C	N	W	C	N	
10. The room temperature was too warm (W) or too cold (C) or neither (N)		x				x			x			x			x	***OU1:*** *"Not warm, actually it was rather cool." "If it gets cooler, it would be disturbing because I"d shiver."* ***OU2:*** *"No."* ***OU3:*** *"Not too warm, not too cold."* ***OU4:*** *"It was quite alright, not too cold, not too warm."* ***OU5:*** *"No."*

Only one operational/untrained rater (OU1) mentioned that the room was *"rather cool"*. He did not clearly state that it was disturbing while he was rating. It would be disturbing only if *"it gets cooler"*. The others concurred that the room was neither too cold nor too warm. So it might be said that this sub-themes did not affect their ratings.

Table 4.57 shows the physical settings if the linguistic/trained raters felt the room was too dark or too lighted or neither.

Table 4.57: Physical settings in terms of the room lighting felt by the linguistic/trained raters (LT)

Sub-themes	Raters															Meaning units
	LT1			LT2			LT3			LT4			LT5			
	D	L	N	D	L	N	D	L	N	D	L	N	D	L	N	
11. The room was too dark (D) or too lighted (L) or neither (N)			x			x			x			x			x	***LT1:*** *"No. It was fine"* ***LT2:*** *"No."* ***LT3:*** *"No, it was comfortable."* ***LT4:*** *"No."* ***LT5:*** *"No. It was okay."*

None of the linguistic/trained raters said that the room they conducted their ratings was too dark or too lighted. It might be said that the room lighting did not affect their ratings.

Table 4.58 shows the physical settings if the linguistic/untrained raters felt the room was too dark or too lighted or neither.

Table 4.58: Physical settings in terms of the room lighting felt by the linguistic/untrained raters (LU)

Sub-themes	Raters															Meaning units
	LU1			LU2			LU3			LU4			LU5			
	D	L	N	D	L	N	D	L	N	D	L	N	D	L	N	
11. The room was too dark (D) or too			x			x			x			x			x	***LU1:*** *"No. It was quite alright."*

lighted (L) or neither (N)	***LU2:*** *"No."* ***LU3:*** *"No. It was fine."* ***LU4:*** *"No. It was fine."* ***LU5:*** *"No."*

All of the linguistic/untrained raters did not have any problem with the room lighting. This factor can be concluded as it did not affect their ratings.

Table 4.59 shows the physical settings if the operational/trained raters felt the room was too dark or too lighted or neither.

Table 4.59: Physical settings in terms of the room lighting felt by the operational/trained raters (OT)

Sub-themes	Raters															Meaning units
	OT1			OT2			OT3			OT4			OT5			
	D	L	N	D	L	N	D	L	N	D	L	N	D	L	N	
11. The room was too dark (D) or too lighted (L) or neither (N)			x			x			x			x			x	***OT1:*** *"No."* ***OT2:*** *"No. Everything was good."* ***OT3:*** *"No, not dark not lighted"* ***OT4:*** *"No."* ***OT5:*** *"No. We were fine."*

None of the raters in the operational/trained category noted the lighting problem. This might suggest that it did not affect their ratings.

Table 4.60 shows the physical settings if the operational/untrained raters felt the room was too dark or too lighted or neither.

Table 4.60: Physical settings in terms of the room lighting felt by the operational/untrained raters (OU)

Sub-themes	Raters															Meaning units
	OU1			OU2			OU3			OU4			OU5			
	D	L	N	D	L	N	D	L	N	D	L	N	D	L	N	
11. The room was too dark (D) or too lighted (L) or neither (N)			x			x		x				x			x	***OU1:*** *"No."* ***OU2:*** *"No."* ***OU3:*** *"It was a little lighted, not much, just a little."* ***OU4:*** *"No."* ***OU5:*** *"No."*

Almost all operational/untrained raters denied that the room was too dark or too lighted. Only OU3 remarked that *"it was a little lighted.* Though it was *"not much, just a little"*, it might somehow affect his rating.

Table 4.61 shows the physical settings if the linguistic/trained raters felt the room was too noisy.

Table 4.61: Physical settings in terms of noise felt by the linguistic/trained raters (LT)

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
12. The room was too noisy		x		x		x		x		x	***LT1:*** *"No. I heard some noise but it wasn"tloud enough to say it was noisy."* ***LT2:*** *"It was quiet." "I heard a phone ring once but it wasn"tannoying. Normally it"snoisier in my office so I'm quite familiar with it."* ***LT3:*** *"No, not at all. It was very quiet. It's even noisier in my office. Someone always chats there."* ***LT4:*** *"No."* ***LT5:*** *"No."*

Not even one linguistic/trained rater complained about the room noise. Two of them (LT1 and LT2) said that they *"heard some noise"* and *"a phone ring"* but *"it wasn"tloud enough to say it was noisy"* and *"it"seven noisier"* in her office. LT3, LT4 and LT5 did not notice any noise at all. Therefore, it might suggest that noise did not affect their ratings.

Table 4.62 shows the physical settings if the linguistic/untrained raters felt the room was too noisy.

Table 4.62: Physical settings in terms of noise felt by the linguistic/untrained raters (LU)

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
12. The room was too noisy		x		x		x		x		x	***LU1:*** *"Sometimes I heard people chatting. It wasn"t annoying but I heard it."* ***LU2:*** *"No. I didn"tnotice any sound. By the way, this pair of headphones was a little too tight. It hurt my ears a bit."* ***LU3:*** *"No. sometimes I heard something but it wasn"t disturbing."* ***LU4:*** *"No."* ***LU5:*** *"No." "I didn"t hear anything, perhaps because I wore headphones."*

None of the linguistic/untrained raters noticed that they were disturbed or annoyed by any sound or noise. LU1 *"heard people chatting"* but *"it wasn"tannoying"*. This sub-themes might be concluded as not having any effect on their ratings. Nonetheless, LU2 complained about the headphones which were *"a little too tight"* that *"it hurt* his *ears a bit"*. This unexpected sub-themes might be somehow disturbing and affected her rating.

Table 4.63 shows the physical settings if the operational/trained raters felt the room was too noisy.

Table 4.63: Physical settings in terms of noise felt by the operational/trained raters (OT)

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
12. The room was too noisy		x		x	x			x		x	***OT1:*** *"No."* ***OT2:*** *"No." "When the headphone was on, everything was quiet but the headphone was too tight. I think you need to have the one that covers the ears, not the one that pushing the ears." "Or it could be my head is big, I don"t know." "But it wasn"t comfortable. You need the one that covers the ears. It"d be better. It"ll lock out the sound completely." "But it doesn"t affect the rating."* ***OT3:*** *"A little bit. I heard somebody talking." "Well, you hear it. You"re distracted and then you miss what they said and then you have to go back. Listen to it again. It"s like ...* (laughter) *... distracting because it"s all over your ears. It"s not like a soundproof area which it just helps when you try to concentrate."* ***OT4:*** *"No. It was alright when I put the headphones on. It was a little noisy when the cleaner did the vacuum. But after I told her to stop, it was fine."* ***OT5:*** *"No, because just two of us."*

OT2"s complaint is the same as LU2 which is about the headphones. *"The headphone was too tight"* and *"it wasn't comfortable."* However, he insisted that *"it doesn"taffect the rating."* OT3 *heard "somebody talking"* and *"distracted"*. It also made him *"miss what they said"*. He had to *"go back"* and *"listen to it again"*. The other three (OT1, OT4 and OT5) did not have any problem with the room noise. OT3 was solely affected by the noise in his rating.

Table 4.64 shows the physical settings if the operational/untrained raters felt the room was too noisy.

Table 4.64: Physical settings in terms of noise felt by the operational/untrained raters (OU)

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
12. The room was too noisy		x		x		x	x			x	***OU1:*** *"No. it was quiet." "If it gets noisy, it"d disturb me."* ***OU2:*** *"No until five p.m. after I finished listening to all three speech samples and took the headphones off I heard some people chatting. It wasn't really annoying but I heard it while I was rating."* ***OU3:*** *"No, just a little. Not noisy."* ***OU4:*** *"There was some interference from people chatting because the headphones were not fully fit. However, if they were, it might hurt." "It affected my rating."* ***OU5:*** *"No. I heard some noise but it wasn"t annoying."*

OU4 obviously stated that he was affected by the noise in his rating. Three raters (OU2, OU3 and OU5) said that hey heard some noise but *"it wasn"tannoying"*. Only OU1 said that the room was quiet. In conclusion, the noise might affect the rating of one rater (OU4).

Table 4.65 shows the preferred place to do the rating of the linguistic/trained raters.

Table 4.65: Preferred rating place of the linguistic/trained raters (LT)

Sub-themes	Raters															Meaning units
	LT1			LT2			LT3			LT4			LT5			
	O	L	P	O	L	P	O	L	P	O	L	P	O	L	P	
13. A preferred place to do the rating i.e. in an office (O), in a sound lab (L) or some places else (P)			x	x					x	x					x	***LT1:*** *"I"d rather go for a place like a coffee shop in a gas station that has a private corner but also has a view for me to see what"s going on around. A place with privacy but not isolated." "If I have to choose between at home and at the office, I prefer the office because there are too many distractions at home such as those TV programs, those drying clothes waiting for ironing."* ***LT2:*** *"It can be anywhere but not so noisy. In an office where people around are chatting is acceptable." "I prefer at the office because it might be too comfortable at home." "The atmosphere in an office is more appropriate to work. As I said, it"s too comfy at home and there is something else to do too."* ***LT3:*** *"Not in a sound lab and, if I can choose, I wouldn"t use this kind of headphones." "I prefer listening from a loudspeaker. That"s the best." "If not a loudspeaker, it*

	may be earphones but not headphones." ***LT4:*** *"I prefer at the office alone in an isolated space."* ***LT5:*** *"A place where it"s set up specially for rating. I mean not for some other purposes such as at home because there are many distractions nor in a park. A quiet place like here."*

When being asked about each linguistic/trained rater"s preferred place to conduct her rating, LT1 said that she preferred *"a place with privacy but not isolated."* This might be said that she did not conduct the rating in the place of her preference. It might have some effect on her rating. LT2, LT4 and LT5 preferred doing it at an office which was the place where they did their ratings. Hence, the place itself should not affect their ratings. LT3 just mentioned that she did not like doing it in a sound lab. She emphasized more on the equipment that she *"prefers listening from a loudspeaker"* to headphones.

Table 4.66 shows the preferred place to do the rating of the linguistic/untrained raters.

Table 4.66: Preferred rating place of the linguistic/untrained raters (LU)

Sub-themes	Raters															Meaning units
	LU1			LU2			LU3			LU4			LU5			
	O	L	P	O	L	P	O	L	P	O	L	P	O	L	P	
13. A preferred place to do the rating i.e. in an office (O),		x				x			x		x		x			***LU1:*** *"I prefer a listening sound lab."* ***LU2:*** *"Either at home or at the office. Just a quiet place." "If I have to choose, I prefer doing it at home perhaps because I was accustomed to it and it has more privacy. At the office people may come around chatting with me." "It*

in a sound lab (L) or some places else (P)	*doesn"t have to be air-conditioned but quiet." "A place with windows so I can look out to see outside."* ***LU3:*** *"I choose to do it at home in my private room where I can relax like lying on a bed. It"s like in a room with privacy where I can do anything I want to."* ***LU4:*** *"In a sound lab should be nice. A lab where doesn"t have any disturbing noise. Here is okay. It"s okay if there"s nobody around but there were people working. Some talked through the phone so I heard some noise. It wasn"t disturbing but it was a little distracting. It wasn"t noisy but a little distracting."* ***LU5:*** *"Something like this. Where I did was okay." "Quiet." "I wouldn"tdo it at home. I prefer at the office." "The important thing is distraction. There wouldn"t be those distracting noise because I need to concentrate to what I listen, especially when I"m unfamiliar with those terms. I had to focus on some other points such as the language usage, their confidence, the use of clauses, etc. So it must be carefully listened. I had to skip those technical terms because I had no idea what they were."*

Some linguistic/untrained raters (LU1 and LU4) mentioned that they preferred doing their ratings in a sound lab. Though LU4 said *"here is okay"* but *"I heard some noise. It wasn"tdisturbing but it was a little distracting"*. LU2 and LU3 preferred doing their ratings at home. Only LU5 said that in the office like the one he did the rating *"was okay"*. In conclusion, the place where they conducted their ratings might have effect on four linguistic/untrained raters (LU1, LU2, LU3 and LU4) because they preferred some other places such as in a sound lab or at home.

Table 4.67 shows the preferred place to do the rating of the operational/trained raters.

Table 4.67: Preferred rating place of the operational/trained raters (OT)

Sub-themes	Raters															Meaning units
	OT1			OT2			OT3			OT4			OT5			
	O	L	P	O	L	P	O	L	P	O	L	P	O	L	P	
13. A preferred place to do the rating i.e. in an office (O), in a sound lab (L) or some places else (P)	x			x			x			x					x	***OT1:*** *"I like a room more to a bright side." "Colder, more than warmer." "Where I was, was fine." "I prefer in the office because at home there is more distraction, situation that you can"t control at home."* ***OT2:*** *"I prefer in the office like this. You have too many distractions at home. So you"d better come to a confined space but I think you need a space not like a small cubical. You could do it on a sofa if you like." "It could be in a place that doesn"t have distractions." "Office would be better, I think."* ***OT3:*** *"At the office but in a secluded area because if I"m at home there"ll be the kids, too many distractions. At the office you know you"re there to work. That"ll be quiet." "Quiet place and not cold." "It doesn"t have to be isolated, just quiet."* ***OT4:*** *"It has the pros and cons doing this at home. It"s like no time limit. There"s no time frame for rating so we can get more in depth. It depends on persons too. Some may dig deeper, some may not. This may not be fair for the test-takers. At an office we have that time frame." "For me, I prefer at the office." "An isolated place. Alone."*

	"Quiet." "It doesn"t have to be a sound lab." ***OT5:*** *"At least have a window. A room with a window. Sometimes, you know, when you get ... rating is so stressful. At least if you could look out the window while you listen to the speech samples. That would ease your stress off and that could ... I don"t know if it would favor the interviewee or not. But at least the stress of the rater would be much lower if he could rest his eyes on something outside ... quite distant." "I guess this room would be better, just to listen and look out and listen."*

OT4 gave some interesting comments about the places to conduct the rating. He pointed out that it might be unfair for test-takers if the rating took place at home because *"there"s no time frame for rating"* so the rater can *" get more in depth"* while doing it *"at an office we have that time frame".* However, he preferred the place like in the office which was the same as OT1, OT2 and OT3. OT5 preferred *"a room with a window",* so he could look out while listening to the speech samples and it would ease his stress off. He thought, *"the stress of the rater would be much lower if he could rest his eyes on something outside".* In conclusion, every rater in this group seemed to be satisfied with the place they did their ratings. So, it might be concluded that the place was not a factor that affected their ratings.

Table 4.68 shows the preferred place to do the rating of the operational/untrained raters.

Table 4.68: Preferred rating place of the operational/untrained raters (OU)

Sub-themes	Raters															Meaning units
	OU1			OU2			OU3			OU4			OU5			
	O	L	P	O	L	P	O	L	P	O	L	P	O	L	P	
13. A preferred place to do the rating i.e. in an office (O), in a sound lab (L) or some places else (P)	x			x			x				x		x			***OU1:*** *"I prefer at home but it might be distracting." "If I have to choose, I'd come here to the office because the setting can be controlled."* ***OU2:*** *"It should be in an office where there is no distraction. If I can choose it should be quiet because it requires lots of concentration."* ***OU3:*** *"I think at an office is better. I like sitting by the window to get a good view of the outside."* ***OU4:*** *"Not at home. I prefer in a sound lab, alone." "It'd be better if it"s a soundproof room like the on at the Institute of Aviation Medicine because it needs lots of concentration. I can even see faces of the candidates in my thought while rating. It has quite an effect." "Yes, it must be quiet. Concentration is the most crucial thing."* ***OU5:*** *"It doesn"tmake any difference. At the office should be okay because it might be distracting at home. It'll be nice to be isolated and quiet."*

Four out of five operational/untrained raters preferred rating at an office. Only one rater (OU4) preferred doing it in a sound lab. Thus, it might be said that he was the only rater who was affected by the place where he did his rating because it was not in a sound lab as his preference, but in an office.

Table 4.69 shows the rating strategies used by the linguistic/trained raters.

Table 4.69: The rating strategies used by the linguistic/trained raters (LT)

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
14.1 „Listening without stopping" strategy		x		x		x		x	x		***LT1:*** *"Not even once. I listened and stopped and took notes."*
	x		x		x		x		x		***LT2:*** *"Not even once."*
14.2 „Listening/stopping/ note-taking" strategy											***LT3:*** *"No."*
											LT4: *"No, I didn"t."*
											LT5: *"Yes, the third guy because it was the shortest. For the first and the second, I kept listening until there was something then I stopped, went back and listened again." "I listened for the third sample without stopping because it was the shortest and probably because he was the last." "I had listened to the first two so I had some idea of the pattern of the test."*

Almost all linguistic/trained raters did not listen to any speech sample from the beginning to the end without stopping before rating, except LT5 who was the sole rater who admitted she did that only when listening to the third sample. It might suggest that most of the raters in this category used the strategy of „listening/stopping/note-taking" before rating their candidates. LT5 employed the same strategy to her first two candidates, though she utilized different strategy of „listening without stopping" to the third test-taker

"*because it was the shortest and probably because he was the last"* after she *"had listened to the first two"* and *"had some idea of the pattern of the test."*

Table 4.70 shows the rating strategies used by the linguistic/untrained raters.

Table 4.70: The rating strategies used by the linguistic/untrained raters (LU)

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
14.1 „Listening without stopping" strategy		x		x		x	x		x		***LU1:*** *"No."* ***LU2:*** *"No, not at all."*
14.2 „Listening/stopping/ note-taking" strategy	x		x		x		x			x	***LU3:*** *"I listened from the beginning to half way and went back to listen from the beginning again, especially the first speech sample because I had no idea what it was so I could familiarize myself with the test tasks. I listened, stopped and went back to listened again when I wanted to make sure of some certain parts."* ***LU4:*** *"Yes, for the second and the third candidates because I got some experience from the first one how the process went on. I didn"t do the same for the first candidate because there were some points which I wasn"t sure especially about the technical terms they used so I had to play back and forth."* ***LU5:*** *"Yes, all three. I didn"tstop any of them. I*

	listened from the beginning to the end and then rated them."

Three linguistic/untrained raters (LU1, LU2 and LU3) accepted that they did not listen to any speech sample from the beginning to the end without stopping before rating. It might be said that these raters in this category used the strategy of „listening/stopping/note-taking" before rating their candidates. LU4 exercised the same strategy to his first candidate. He switched to the different strategy when listening to the other two by listening to those speech samples from the beginning to the end without stopping before rating. LU5 was the only rater in this group who exploited the „listening without stopping" strategy to all his three candidates.

Table 4.71 shows the rating strategies used by the operational/trained raters.

Table 4.71: The rating strategies used by the operational/trained raters (OT)

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
14.1 „Listening without stopping" strategy	x			x		x	x		x		***OT1:*** *"Yes, just the first one. The second one I had to stop twice, just to write down a note. Momentarily, just to complete my thoughts."*
14.2 „Listening/stopping/ note-taking" strategy	x		x		x						***OT2:*** *"No." "I only went back for the parts I told you that I didn"t expect the answer but it was mostly divided into three parts except for the last."*
											OT3: *"No. I stopped at certain parts."*

	OT4: *"Yes, for all three samples." "I listened to and took notes when I noticed something. Then I went back and listened to those parts again after finishing the whole sample."* ***OT5:*** *"Yeah. I listened to the whole thing first and then remembered where each part of the test is and then skimmed through that part." "I listened to the whole thing then came back to the specific parts which I think affect the rating."*

OT1 applied two different strategies to his subjects. He used the „listening without stopping" strategy to the first candidate but used the „listening/stopping/note-taking" strategy when listening to the other two. OT2 and OT3 strictly utilized the „listening/stopping/note-taking" strategy for all three test-takers. OT4 and OT5 applied the same strategy by „listening/stopping/note-taking" first and then *"went back and listened to those parts again after finishing the whole sample"* (OT4) and *"listened to the whole thing then came back to the specific parts which I think affect the rating"* (OT5).

Table 4.72 shows the rating strategies used by the operational/untrained raters.

Table 4.72: The rating strategies used by the operational/untrained raters (OU)

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
14.1 „Listening without stopping" strategy		x	x			x	x		x		***OU1:*** *"No."* ***OU2:*** *"Yes. I listened just once from the beginning to the end for all three samples."*
14.2 „Listening/stopping/	x			x	x			x		x	

note-taking" strategy	***OU3:*** *"No. I stopped periodically."* ***OU4:*** *"Yes. I didn"t stop at all while listening to all three samples."* ***OU5:*** *"Yes, all three samples."*

OU1 and OU3 utilized the „listening/stopping/note-taking" strategy for all three test-takers while OU2, OU4 and OU5 applied the „listening without stopping" strategy**.**

Table 4.73 shows the number of times of listening before rating of the linguistic/trained raters.

Table 4.73: The number of times of listening before rating of the linguistic/trained raters (LT)

Sub-themes	**Rater LT1**		**Rater LT2**		**Rater LT3**		**Rater LT4**		**Rater LT5**		**Meaning units**
	Once	**>1**	**Once**	**>1**	**Once**	**>1**	**Once**	**>1**	**Once**	**>1**	
15. Times of listening before rating		x		x	x			x		x	***LT1:*** *"I listened from the beginning, stopped and played backward to listen to either when he said something good or something which I didn"tunderstand or when he said something totally wrong, then I took notes."* ***LT2:*** *"I went backward to listen again if I felt there was something wrong." "Not so often."* ***LT3:*** *"Just once. I kept listening and stopping when I wanted to concentrate on some certain parts." "If I*

	have time, I may listen from the beginning to the end but not today." "I must really have time to do that but I seldom do that." "It depends on the number of samples too." ***LT4:*** *"I went back and forth but not so often as in the phraseology part, there wasn"t much to rate except pronunciation and listening comprehension because they answered in standard phraseology."* ***LT5:*** *"Many times."*

The only linguistic/trained rater who admitted that she listened to the speech samples just once before rating was LT3. She added that she might have listened more than once if she had time. It also "*depends on the number of samples too."* The others listened to their samples more than once.

Table 4.74 shows the number of times of listening before rating of the linguistic/untrained raters.

Table 4.74: The number of times of listening before rating of the linguistic/untrained raters (LU)

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Once	>1	Once	>1	Once	>1	Once	>1	Once	>1	
15. Times of listening before rating		x		x		x	x	x	x		***LU1:*** *"I repeatedly listened to the first candidate many times trying to perceive each part because I didn"t understand the situations."* ***LU2:*** *"It was three or four times for the first candidate because I didn"t get what he said so I had to go*

		backward to listen again. A few times for the second candidate and just once or twice for the third because his sample was short and I'd also got acquainted with the content." ***LU3:*** *"Many times but I didn"t listen to the whole speech. Just back and forth to listen to some specific parts."* ***LU4:*** *"I listened to the first candidate and stopped, played back and listened again for quite a few times because I didn"t understand the context they were talking. But after getting the pictures I listened to the second and the third candidates just once."* ***LU5:*** *"Just once for each sample."*

LU5 was the only linguistic/untrained rater who accepted that he listened to all of his samples *"just once for each sample."* LU4 did it once *"after getting the pictures"* from the first candidate. She *"listened to the second and the third candidates just once."* The others listened to them more than once.

Table 4.75 shows the number of times of listening before rating of the operational/trained raters.

Table 4.75: The number of times of listening before rating of the operational/trained raters (OT)

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Once	>1	Once	>1	Once	>1	Once	>1	Once	>1	
15. Times of	x	x		x		x		x		x	***OT1:*** *"Two times for the first one and then a short*

listening before rating	*clip." "The second one and the third one – just once."* ***OT2:*** *"I listened many times but I didn"t listen to all of it. I listened to parts of it."* ***OT3:*** *"Two or three times, just to make sure I understood more exactly. Just a few times."* ***OT4:*** *"I listened from the beginning to the end once. While listening I took notes of some particular parts by jotting down the time they occurred. After finishing I went back to those parts and listened to them again."* ***OT5:*** *"Well, I couldn"t say twice because I didn"t listen to the whole thing the second time again. So I listened to it once and went through the parts where I thought he made mistakes then assessed on that." "So it"s sort oftwice."*

OT1 stated that he listened to the first speech sample twice but just once for the second and the third. The other operational/trained raters listened to their subjects more than once before rating.

Table 4.76 shows the number of times of listening before rating of the operational/untrained raters.

Table 4.76: The number of times of listening before rating of the operational/untrained raters (OU)

Sub-themes	Rater OU1		Rater OU1		Rater OU1		Rater OU1		Rater OU1		Meaning units
	Once	>1	Once	>1	Once	>1	Once	>1	Once	>1	
15. Times of listening before rating		x	x			x	x		x		***OU1:*** *"I listened from the beginning to the end and took notes whenever I wanted. I went backward a few times to listen to some parts."* ***OU2:*** *"Just once."* ***OU3:*** *"A few times."* ***OU4:*** *"Just once."* ***OU5:*** *"Once for each candidate."*

OU1 and OU3 said that they listened to their speech samples more than once. The other operational/untrained raters (OU2, OU4 and OU5) listened to them *"just once."*

Table 4.77 shows the rating strategy of note taking used by the linguistic/trained raters.

Table 4.77: The rating strategy of note taking used by the linguistic/trained raters (LT)

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
16. Note taking	x		x		x		x		x		***LT1:*** *"Very often." "Two pages for each test takers."* ***LT2:*** *"Always."*

	LT3: *"I took notes from time to time. Quite frequently."* ***LT4:*** *"Yes. All the times."* ***LT5:*** *"Frequently."*

All linguistic/trained raters accepted that they took notes *"very often"* (LT1), *"always"* (LT2), *"frequently"* (LT3 and LT5) or *"all the times"* (LT4).

Table 4.78 shows the rating strategy of note taking used by the linguistic/untrained raters.

Table 4.78: The rating strategy of note taking used by the linguistic/untrained raters (LU)

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
16. Note taking	x		x		x		x		x		***LU1:*** *"Yes. Very often."* ***LU2:*** *"Yes, almost all the time."* ***LU3:*** *"Yes but not so often, just sometimes." "I took notes when they were obvious. For example the pronunciation, if I understood what they spoke I kept going on. I took notes when some difficulties arose."* ***LU4:*** *"Yes. It was frequently for the first candidate but only sometimes for the latter two because the context was new to me." "After I got more familiar with the second and the third I took*

											notes in my mind." ***LU5:*** *"All the times."*

LU1, LU2 and LU5 said that they took notes *"very often"* or *"all the times"*. LU3 said that she took notes less often, just only *"when they were obvious"* and *"when some difficulties arose"*. LU4 took notes *"frequently for the first candidate"* but *"only sometimes for the latter two"*.

Table 4.79 shows the rating strategy of note taking used by the operational/trained raters.

Table 4.79: The rating strategy of note taking used by the operational/trained raters (OT)

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
16. Note taking	x		x		x	x	x		x		***OT1:*** *"Every subject, every single part of the interview." "Both positive and negative."* ***OT2:*** *"Often. That"s the way to do the rating if you can"t completely remember. "* ***OT3:*** *"If I do a difficult one, I took quite a few notes. But if it"san easy one, not so many notes because I listened to it as a whole." "For sample number one I wrote a lot of notes but then you find out that you got distracted by all the notes. So in the end I wrote less and less. I thought … you know … I listened to it as a whole. Number two was less because number two was not so good and number three was really good so I thought I need it less so I just listened to the whole." "I changed my strategy from number one, number two and number three." "I was listening just too much to*

	the certain parts instead of listening to the whole. So I changed my style." ***OT4:*** *"Frequently."* ***OT5:*** *"As the tape went along. Yeah, quite often." "Sometimes I remembered the mistakes he made because ... as you rate more and more you could remember ... it"s a specific pattern that Thai people make." "Whether it"s the „r," „l," the plural „s", singular „s" or whatever. It"s basically with all. The people who get a „four," a „thræ" ... I mean as you gain more experience, you don"t have to take that much notes."*

Four out of five operational/trained raters accepted that they took notes *"quite often"* (OT2 and OT5) or *"frequently"* (OT4). OT1 stressed that he even took notes for *"every subject, every single part of the interview" "both positive and negative".* OT2 also added that *"that"s the way to do the rating if you can"t completely remember".* However, OT3 had a different point of view. He thought that *"you got distracted by all the notes".* That was the reason why he changed his strategy from the speech sample number one who he *"wrote a lot of notes"* to number two and number three who he *"just listened to the whole."*

Table 4.80 shows the rating strategy of note taking used by the operational/untrained raters.

Table 4.80: The rating strategy of note taking used by the operational/untrained raters (OU)

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
16. Note taking	x			x	x		x			x	***OU1:*** *"From time to time."*

	OU2: *"No."* ***OU3:*** *"Yes, frequently."* ***OU4:*** *"Not so often, only when I found errors."* ***OU5:*** *"I didn"t take note at all but I roughly gave the scores and might change them if I found something else."*

The operational/untrained group is the only group which has raters who did not take notes at all. They were OU2 and OU5. The other two raters in this batch just did it *"from time to time"* (OU1) and *"not so often"* (OU4). The sole rater who took notes *"frequently"* was OU3.

Table 4.81 shows the rating strategy of tape stopping used by the linguistic/trained raters.

Table 4.81: The rating strategy of tape stopping used by the linguistic/trained raters (LT)

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
17. Tape stopping (other than to take notes)											***LT1:*** *"Very often."* ***LT2:*** *"Just once. To go to the toilet."* ***LT3:*** *"Once, to go to the toilet."* ***LT4:*** *"Twice. Once to answer the phone and the other to go to the toilet."*

	LT5: *"I stopped four or five times to answer phone calls."*

Every linguistic/trained rater stopped the speech sample tape at least once for two different reasons - going to the toilet (LT2 and LT3) or answering the phone (LT5) or both (LT4). LT1 did not state the reason of her stopping but she said that she stopped *"very often"*.

Table 4.82 shows the rating strategy of tape stopping used by the linguistic/untrained raters.

Table 4.82: The rating strategy of tape stopping used by the linguistic/untrained raters (LU)

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
17. Tape stopping (other than to take notes)	x		x			x	x		x		***LU1:*** *"Not so often, just a few times for each candidate."* ***LU2:*** *"Yes. Quite often for the first one. Less and less for the other two."* ***LU3:****"No. I didn"t stop for any other reason. I just stopped to take notes."* ***LU4:*** *"No, I didn"tgo to the toilet at all but I stopped to answer the phone twice, once during the first rating and another during the third."* ***LU5:*** *"Once because of a technical problem to set up the system, then I went to the restroom."*

LU3 was the only linguistic/untrained rater who did not stop the tape to answer the phone or to go to the toilet. The others stopped the tape for the most two common reasons - going to the toilet (LU5) or answering the phone (LU4). LU1 and LU2 did not state the reason of their stopping.

Table 4.83 shows the rating strategy of tape stopping used by the operational/trained raters.

Table 4.83: The rating strategy of tape stopping used by the operational/trained raters (OT)

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
17. Tape stopping (other than to take notes)	x		x			x	x		x		***OT1:*** *"Rare. Once or twice. Once with number two and two with number three and also number one, I think.* ***OT2:*** *"Quite often."* ***OT3:*** *"No, I didn"t go to the toilet or picked up any phone call."* ***OT4:*** *"I went to the toilet twice."* ***OT5:*** *"After listened to the whole thing I went to the men"s room once." "I finished the first two then I took a break." "So I didn"t stop in between."*

OT3 was the only rater in the operational/trained group who said that he *"didn"tgo to the toilet or picked up any phone call"*. OT4 and OT5 stopped the tape to go to the toilet. Even so, OT5 added that he *"didn"tstop in between"*. He went there *"after listened to the whole thing"*. OT1 and OT2 did not state the reason of their stopping.

Table 4.84 shows the rating strategy of tape stopping used by the operational/untrained raters.

Table 4.84: The rating strategy of tape stopping used by the operational/untrained raters (OU)

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
17. Tape stopping (other than to take notes)		x		x	x		x			x	***OU1:*** *"I stopped the tape just to take notes." "I didn"t take any brake."* ***OU2:*** *"No."* ***OU3:*** *"Yes. I went to the toilet once after finishing the first two samples."* ***OU4:*** *"I went to the toilet once after finishing the first two."* ***OU5:*** *"No, not even once."*

Three operational/untrained raters (OU2 and OU5) stated that they did not stop the tape for any reason. OU3 and OU4 said that they stopped to go to the toilet. OU1 *"stopped the tape just to take notes"*. He *"didn"t take any brake"*.

Table 4.85 shows the rating strategy of stopping the tapes to listen for certain parts used by the linguistic/trained raters.

Table 4.85: The rating strategy of stopping the tapes to listen for certain parts used by the linguistic/trained raters (LT)

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
18. Stopping the tapes to listen for certain parts	x		x		x		x		x		***LT1:*** *"Very often."* ***LT2:*** *"I did when I either wasn"tsure or couldn"tget what they spoke."* ***LT3:*** *"I stopped the tape quite frequently because I couldn"t catch it. For example, if there was something I thought I should have jotted down and while I was taking it that part had passed, so I had to rewind the tape and listen again." "Sometimes I had to do this a few times before I finished with that part."* ***LT4:*** *"Sometimes."* ***LT5:*** *"Yes and I went to the toilet once."*

All linguistic/trained raters admitted that they stopped to listen for the certain parts of the speech samples.

Table 4.86 shows the rating strategy of stopping the tapes to listen for certain parts used by the linguistic/untrained raters.

Table 4.86: The rating strategy of stopping the tapes to listen for certain parts used by the linguistic/untrained raters (LU)

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
18. Stopping the tapes to listen for certain parts	x		x		x		x	x		x	***LU1:*** *"I stopped to listen to the second candidate more often than the first and the third ones."* ***LU2:*** *"Quite frequent for the first candidate because I didn"t clearly get the questions. Less frequent for the other two because I got more familiar with the questions."* ***LU3:*** *"Yes. Sometimes."* ***LU4:*** *"Yes, for the first candidate but none for the other two."* ***LU5:*** *"Not at all."*

Almost all raters in the linguistic/untrained rater category said that they stopped to listen for the certain parts of the speech samples, except LU5 who said that he did not do that *"at all"*. LU4 was the only rater who used the mixed strategy. He did it *"for the first candidate but none for the other two"*.

Table 4.87 shows the rating strategy of stopping the tapes to listen for certain parts used by the operational/trained raters.

Table 4.87: The rating strategy of stopping the tapes to listen for certain parts used by the operational/trained raters (OT)

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
18. Stopping the tapes to listen for certain parts	x		x		x		x		x		***OT1:*** *"For the first person, I had twice. The second I did it one time and the third I did it twice."* ***OT2:*** *"Many times."* ***OT3:*** *"Just a few times."* ***OT4:*** *"Frequently."* ***OT5:*** *"Certain parts? Not so often because, like I said, after you interview for a while you know the common mistakes that people make. I mean the common mistakes that Thai people make. I would say Thai people make. If I was interviewing some other nationality, he would have a mistake somewhere else." "Just sometimes. If you hear their mistakes, there"s no use stopping. Just write the comment and listen to the next part at the same time."*

All operational/trained raters admitted that they stopped to listen for the certain parts of the speech samples. Some (OT2 and OT4) did it *"frequently"*. Some (OT2, OT4 and OT5) did it *"just a few times"*.

Table 4.88 shows the rating strategy of stopping the tapes to listen for certain parts used by the operational/untrained raters.

Table 4.88: The rating strategy of stopping the tapes to listen for certain parts used by the operational/untrained raters (OU)

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
18. Stopping the tapes to listen for certain parts	x			x	x			x		x	***OU1:*** *"Around seven to ten times per candidate."* ***OU2:*** *"No."* ***OU3:*** *"Yes. I stopped a few times for each sample."* ***OU4:*** *"Never. I listened just once."* ***OU5:*** *"Not at all."*

The majority of the operational/untrained raters (OU2, OU4 and OU5) did not stop to listen for any certain part of the speech samples. OU3 stopped just *"a few times for each sample"* while OU1 stopped more often (*"around seven to ten times per candidate"*).

Table 4.89 shows the rating strategy of concentration on language or content or both used by the linguistic/trained raters.

Table 4.89: The rating strategy of concentration on language or content or both used by the linguistic/trained raters (LT)

Sub-themes	Rater															Meaning units
	LT1			LT2			LT3			LT4			LT5			
	L	C	B	L	C	B	L	C	B	L	C	B	L	C	B	
19. Concentration on language (L)	x			x					x	x			x			***LT1:*** *"Language first."*

or content (C) or both (B)	***LT2:*** *"I concentrated more on the language."* ***LT3:*** *"Fifty-fifty." "The content, in this case, means relevance, if it"s straight to the point or not, not in terms of job-specific aspects."* ***LT4:*** *"More on the language."* ***LT5:*** *"Language."*

Most of the linguistic/trained raters agreed that they concentrated first on the language in their rating. Only LT3 said that she weighted the language and the content equally in her rating.

Table 4.90 shows the rating strategy of concentration on language or content or both used by the linguistic/untrained raters.

Table 4.90: The rating strategy of concentration on language or content or both used by the linguistic/untrained raters (LU)

Sub-themes	Rater															Meaning units
	LU1			LU2			LU3			LU4			LU5			
	L	C	B	L	C	B	L	C	B	L	C	B	L	C	B	
19. Concentration on language (L) or content (C) or both (B)		x		x	x		x					x	x			***LU1:*** *"Content, especially in the interview, in the part that the interviewer asked the question about the advantage and disadvantage of the technology because I regard the part of simulation as the technical term usage which they used phrases or terms that were standardized, not the full forms of language." "All three, especially the latter two, had problems with the last part because they were unable to elaborate their answers. They were just short answers."*

LU2: *"I focused on the structure while listening to the first candidate. Then I realized that I should have concentrated on the content too when listening to the second and the third." "The second candidate answered the questions with short answers while the first candidate"s answers were longer that"s why he made more mistakes with tenses but it didn"t interfere with the meanings." "First I focused more on the language then I focused more on the content. For example when they tried to describe the bomb, the second candidate tried to use some technical terms such as a cylinder. I started to consider if it made any sense to use that word."*

LU3: *"The language."*

LU4: *"I concentrated on both the content and the language."*

LU5: *"It was certainly at the language because I knew nothing about the content."*

Three linguistic/untrained raters (LU2, LU3 and LU5) said that they concentrated on the language while rating. LU4 said that she "*concentrated on both"*. LU1 was the only rater who admitted that she concentrated on the content. It is worth noting that LU5 concentrated on the language because she *"knew nothing about the content"*. LU2 was the one who *"first focused more on the language then focused more on the content* because she *"focused on the structure while listening to the first candidate. Then* she *realized that* she *should have concentrated on the content too when listening to the second and the third"*.

Table 4.91 shows the rating strategy of concentration on language or content or both used by the operational/trained raters.

Table 4.91: The rating strategy of concentration on language or content or both used by the operational/trained raters (OT)

Sub-themes	Rater															Meaning units
	OT1			OT2			OT3			OT4			OT5			
	L	C	B	L	C	B	L	C	B	L	C	B	L	C	B	
19. Concentration on language (L) or content (C) or both (B)	x			x			x			x			x			***OT1:*** *"The language first. "Cause the first part is pronunciation, so I concentrate on how the pronunciation is. But then as soon as you get a feeling, maybe after part one, part two, then you get a feeling of the pronunciation level that you know the level, then I concentrate on the content."* ***OT2:*** *"For this I know that the language was the most important thing. Content was a secondary thing. I can definitely say I concentrate on the language first."* ***OT3:*** *"Language."* ***OT4:*** *"The language."* ***OT5:*** *"I concentrated on the language first. I mean content really doesn"t mean much when you"re rating just the proficiency in English, right? It has something to do with the comprehension, right? But, as you can see, comprehension is the last part, almost the last part. When you rate this guy, you"drate him on his language, on his pronunciation, structure, vocabulary, his fluency, then the comprehension and interactions, right? So I would listen to the language first. "*

All operational/trained raters concentrated on the language while rating. OT2 emphasized that he knew *"that the language was the most important thing. Content was a secondary thing"*.

Table 4.92 shows the rating strategy of concentration on language or content or both used by the operational/untrained raters.

Table 4.92: The rating strategy of concentration on language or content or both used by the operational/untrained raters (LT)

Sub-themes	Rater															Meaning units
	OU1			OU2			OU3			OU4			OU5			
	L	C	B	L	C	B	L	C	B	L	C	B	L	C	B	
19. Concentration on language (L) or content (C) or both (B)			x			x	x					x	x			***OU1:*** *"The tasks did not require much explanation of ideas. Most of them just asked for yes/no answers or patterned sentences. The candidates knew what to answer, what sort of grammar and vocabulary needed." "I focused just on the structure and vocabulary." "Let"s sayI concentrated on both."* ***OU2:*** *"Both." "It"s like watching a movie. You"ve got to look at many aspects. I see if I could understand them then see what level they should be in. So it doesn"t mean that you"re native speaker then you"ve got to be level six."* ***OU3:*** *"I put more weight on the language. I pay attention to the content too but, as I said, I was interested more on the interview. I focused more on the language."* ***OU4:*** *"Both."*

OU5: *"Language"*

Three raters in the operational/untrained batch (OU1, OU2 and OU4) said that they concentrated on both the language and the content while rating. The other two (OU3 and OU5) said that they *"put more weight on the language"*.

Table 4.93 shows the rating strategy of focusing on accuracy or fluency or both used by the linguistic/trained raters.

Table 4.93: The rating strategy of concentration on language or content or both used by the linguistic/trained raters (LT)

Sub-themes	Rater															Meaning units
	LT1			LT2			LT3			LT4			LT5			
	A	F	B	A	F	B	A	F	B	A	F	B	A	F	B	
20. Focus on accuracy (A) or fluency (F) or both (B)	x					x			x	x					x	***LT1:*** *"More on accuracy."* ***LT2:*** *"I focused on both. They come together."* ***LT3:*** *"Both, equally."* ***LT4:*** *"Fluency may be a big part but accuracy is a little more important because this kind of interview assessment is to elicit the candidates" proficiency in using the language so their fluency might not be as good as when they told their own stories. By paraphrasing or explaining the events, they"d certainly lose some of their fluency. But their accuracy would show what their proficiency was." "But it doesn"t mean that their fluency was so bad that it was incomprehensible."* ***LT5:*** *"Both."*

Three out of five linguistic/trained raters (LT2, LT3 and LT5) admitted that they focused on both accuracy and fluency in their ratings. The other two (LT1 and LT4) said that they focused more on accuracy. However, *"it doesn"tmean that their fluency was so bad that it was incomprehensible."* (LT4).

Table 4.94 shows the rating strategy of focusing on accuracy or fluency or both used by the linguistic/untrained raters.

Table 4.94: The rating strategy of concentration on language or content or both used by the linguistic/untrained raters (LU)

Sub-themes	Rater															Meaning units
	LU1			LU2			LU3			LU4			LU5			
	A	F	B	A	F	B	A	F	B	A	F	B	A	F	B	
20. Focus on accuracy (A) or fluency (F) or both (B)			x		x				x	x				x		***LU1:*** *"Which part do you mean?" "For overall I focused on both."* ***LU2:*** *"Fluency because I think, as pilots who are non-native speakers of English, it would be difficult for them to use the exact words to get accuracy. Under some circumstances, it might be too late if they try to get the correct terms." "Well, I think I focus on both but put more weight on fluency."* ***LU3:*** *"Both."* ***LU4:*** *"Accuracy or fluency? Umm ... it"s hard to answer. Fluency may be a matter of each individual style. Or perhaps it may be a normal procedure for the job, especially for pilot-air traffic controller communication. It may require them to speak slowly. Right? So I focused on accuracy more than fluency." "But if you ask if fluency is good, it is if you can." "But today I focused on accuracy." "Fluency is about*

	expectation. If it is expected by the working environment that you should have fluency, then you should have it. It depends on what is required. But in my today rating I focused first on accuracy." ***LU5:*** *"It should have been fluency." "I didn"tfocus much on grammar." "I think that this kind of communication focuses on understanding but understanding will occur when you"re fluent in what you"resaying. I think it"shard for both things to go together, fluency and accuracy. We consider accuracy in terms of using correct terms, correct vocabulary, using specific words for specific terms but not accuracy in terms of creating sentences with correct grammar. It"s not the main Sub-themes in this kind of assessment. So fluency should come first."*

The linguistic/untrained had split ideas. Two of them (LU1 and LU3) stated that they focused on both accuracy and fluency while the other two (LU2 and LU5) accepted that they focused more on fluency and the only one who *"focused first on accuracy"* was LU4.

Table 4.95 shows the rating strategy of focusing on accuracy or fluency or both used by the operational/trained raters.

Table 4.95: The rating strategy of concentration on language or content or both used by the operational/trained raters (OT)

Sub-themes	Rater															Meaning units
	OT1			OT2			OT3			OT4			OT5			
	A	F	B	A	F	B	A	F	B	A	F	B	A	F	B	
20. Focus on			x			x		x			x			x		***OT1:*** *"Both."*

accuracy (A) or fluency (F) or both (B)	***OT2:*** *"I did both because I think both are parts of the criteria."* ***OT3:*** *"First I was concentrating more on the accuracy and number two, number three I went for more on fluency." "Number one was more on accuracy. Number two and number three I went more on fluency because sometimes, okay, they can make mistakes in certain parts so accuracy's not that good. But when you listen to it as a whole, it's not that bad." "Because I was thinking „Oh! He made a mistake". He didn't put that „and". He didn't do this. He didn't do that. And then in the end when you listened to it, you think he meets the requirement for this level. First I thought he didn't meet it, I looked at it again ... well ... it says if he can do this then okay he's in this level." "But then in the end I listened to it just for fluency and overall ... comprehension ... to make decision."* ***OT4:*** *"Accuracy."* ***OT5:*** *"Um ... fluency. More on fluency. I mean at least the answer has to be right ... according to the question being asked, right? If it's somewhere along that line then ... I'd say it's fine."*

Two operational/trained raters (OT1 and OT2) said that they focused on both accuracy and fluency. The two raters in this group who focused more on fluency were OT3 and OT5. OT4 was the only rater who insisted that he focused more on accuracy.

Table 4.96 shows the rating strategy of focusing on accuracy or fluency or both used by the operational/untrained raters.

Table 4.96: The rating strategy of concentration on language or content or both used by the operational/untrained raters (OU)

Sub-themes	Rater															Meaning units
	OU1			OU2			OU3			OU4			OU5			
	A	F	B	A	F	B	A	F	B	A	F	B	A	F	B	
20. Focus on accuracy (A) or fluency (F) or both (B)	x				x		x			x					x	***OU1:*** *"I focused more on accuracy." "Because if it"s not accurate, it may lead to miscommunication."* ***OU2:*** *"I focused more on fluency but the overall was that it must be comprehensible."* ***OU3:*** *"I think I focused more on the accuracy. As I always wrote, I was concerned about the articulation. I personally think that it"s important. Even though your language is not good but if you are able to articulate, you can get your point across." "Fluency may sound nice but it"s less important."* ***OU4:*** *"More on accuracy, less on fluency."* ***OU5:*** *"Both."*

The majority of the operational/untrained raters (OU1, OU3 and OU4) said that they focused more on accuracy while one rater (OU2) accepted that he focused more on fluency under the condition that *"the overall was that it must be comprehensible."* OU5 was the only rater in this batch who focused on both accuracy and fluency.

Table 4.97 shows the rating strategy of rating each criterion before or after the overall performance used by the linguistic/trained raters.

Table 4.97: The rating strategy of rating each criterion before or after the overall performance used by the linguistic/trained raters (LT)

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	B	A	B	A	B	A	B	A	B	A	
21. Rating each criterion before (B) or after (A) the overall performance	x		x		x		x		x	x	***LT1:*** *"I had an overall picture of the test takers while I was listening and taking notes." "I rated each criterion first." "For example, this guy who got „four" in most criteria, I would give him „four" instead of „three" if I knew nothing about the ICAO requirement that the overall score must be the lowest score from any criterion."* ***LT2:*** *"I rated each criterion first."* ***LT3:*** *"I rated each criterion first. For me, it depends on the way I listen to the samples. As I said, if I have to rate ten samples in three days, I have time so I'll listen from the beginning to the end once. By doing that, I'll have an overall impression. Then I'll consider each criterion." "In this case, it was each criterion first because while I was jotting I had the picture of each criterion."* ***LT4:*** *"I looked at their overall performance first. For example when I initially listened to the second candidate, I couldn't help feeling that this guy was rather weak, especially in the standard phraseology part. He might say it right but it wasn't so good. He did even worse when describing the object on the runway. But when it comes to the face-to-face conversation, I felt that he was able to cope with the on-the-spot problems though he needed to repeat, to confirm or to clarify but he managed to do it. This was his overall. Then I looked at each area which he was somewhat weak." "I rated each criterion first before giving the overall score. For example, the second candidate, the first guy certainly passed while the third surely failed, this guy was in between. After finish listening I felt that this guy was*

	able to survive. He was in operational level but at the very threshold because his weak point was his vocabulary, his grammar wasn"tgood and his pronunciation was bad. Then I looked at my notes, I felt that no matter how bad he was, he had some other kind of strategies to use such as when he said some words which was hard to understand, he was able to paraphrase that." "I rated each criterion first before giving the overall score because I already knew that was the ICAO requirement." If there is no such requirement, I would rate the overall first before rating each criterion." ***LT5:*** *"Both. Sometimes I had an assumption that he should be at that level by his overall performance but sometimes I couldn"tsee if he was really at that level so I had to look at each criterion." "I rated the first guy"soverall first. For the second guy, I rated each criterion first. I did the same for the third as the first one." "I did differently for the second guy because I felt that he made more mistakes than the first and the third. That"swhy I looked at each separate criterion first while the first and the third seemed not to have much problem. The second guy seemed to have many problems so I looked separately and gave the final score later."*

Almost all linguistic/trained raters (LT1, LT2, LT3 and LT4) said that hey rated each criterion first before rating the overall performance. However, LT1 and LT4 admitted that they *"had an overall picture of the test takers while listening and taking notes"* (LT1) and *"looked at their overall performance first"* (LT4). LT5 was the only rater in this category who used mixed strategies by *"rated the first guy"soverall first"* but *"for the second guy, I rated each criterion first"*. Then she *"did the same for the third as the first one."*

Table 4.98 shows the rating strategy of rating each criterion before or after the overall performance used by the linguistic/untrained raters.

Table 4.98: The rating strategy of rating each criterion before or after the overall performance used by the linguistic/untrained raters (LU)

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	B	A	B	A	B	A	B	A	B	A	
21. Rating each criterion before (B) or after (A) the overall performance	x		x		x		x		x		***LU1:*** *"I rated each criterion first, then the overall."* ***LU2:*** *"I rated each criterion first, then the overall."* ***LU3:*** *"I rated each criterion first, and rated the overall later."* ***LU4:*** *"I rated each criterion first then rated the overall later."* ***LU5:*** *"I rated each criterion first."*

The strategy that the linguistic/untrained raters used in this perspective was unanimous. All of them said that they rated each criterion first before rating the overall performance.

Table 4.99 shows the rating strategy of rating each criterion before or after the overall performance used by the operational/trained raters.

Table 4.99: The rating strategy of rating each criterion before or after the overall performance used by the operational/trained raters (OT)

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	B	A	B	A	B	A	B	A	B	A	
21. Rating each criterion before (B) or after (A) the overall performance	x		x		x		x		x		***OT1:*** *"I rated each subject first, and then rated overall."* ***OT2:*** *"I did each criterion first and overall was the last thing I did."* ***OT3:*** *"Each individual and then the lowest one would be … yeah."* ***OT4:*** *"I rated each criterion first then the overall."* ***OT5:*** *"I rated each criterion first then I rated the overall score."*

The operational/trained raters agreed that all of them rated each criterion first before rating the overall performance.

Table 4.100 shows the rating strategy of rating each criterion before or after the overall performance used by the operational/untrained raters.

Table 4.100: The rating strategy of rating each criterion before or after the overall performance used by the operational/untrained raters (OU)

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	B	A	B	A	B	A	B	A	B	A	
21. Rating each criterion before (B) or after (A) the overall performance	x			x	x			x	x		***OU1:*** *"I rated each criterion first then rated the overall."* ***OU2:*** *"I gave the overall scores first. It would become the big picture in my head. Then I looked at each individual criterion."* ***OU3:*** *"I rated each criterion first."* ***OU4:*** *"I rated the overall performance first then I looked back and rated each criterion."* ***OU5:*** *"I rated each criterion first and overall later."*

Three operational/untrained raters (OU1, OU3 and OU5) said that they rated each criterion first before rating the overall performance while the other two (OU2 and OU4) did it the other way around. They rated the overall performance first.

Table 4.101 shows the rating strategy of concentration on errors used by the linguistic/trained raters.

Table 4.101: The rating strategy of concentration on errors used by the linguistic/trained raters (LT)

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Much	Not much	Much	Not much	Much	Not much	Much	Not much	Much	Not much	
22. Concentration on errors	x			x		x	x	x		x	***LT1:*** *"All the time."* ***LT2:*** *"I kept listening and stopped when there were errors. So it depends on how often they made mistakes. It happened when I was struck by any doubt."* ***LT3:*** *"As necessary when it interfered with the meaning." "Quite often when it interfered with the meaning." "If it doesn"t, I may overlook it." "I put meanings as the main concern."* ***LT4:*** *"Frequently but I didn"t concentrate on the very details such as „oh! you said this without „s" but I rather listened to … what … when I listened I knew that he knew the basic. He might miss them because of some factors such as he didn"t use them quite often or this was the way he was familiar." "I accepted that this was his level. If he was level five*

	or level six, I would concentrate more on his errors." ***LT5:*** *"Four out of five. It means that when he made the same mistakes repeatedly I didn"t have to concentrate on those errors any more." "Let"s say,,sometimes"."*

LT1 said that she concentrated on the errors made by the candidates *"all the time"*. LT2 did that depending on *"how often they made mistakes"* while LT3 paid her attention to the error *"when it interfered with the meaning"*. LT4 *"didn"tconcentrate on the very details"* but she *"would concentrate more on his errors if he was level five or level six"*. LT5 was more or less similar to LT2 in the way that she focused on the errors *"when he made the same mistakes repeatedly"*.

Table 4.102 shows the rating strategy of concentration on errors used by the linguistic/untrained raters.

Table 4.102: The rating strategy of concentration on errors used by the linguistic/untrained raters (LU)

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Much	Not much	Much	Not much	Much	Not much	Much	Not much	Much	Not much	
22. Concentration on errors		x		x		x	x			x	***LU1:*** *"Not so often."* ***LU2:*** *"I concentrated a lot with the first candidate then I had a second thought that it shouldn"t have been that much. I should*

have focused more on the overall if I could understand them. Then I asked myself „do I understand them because I'm Thai?" because they were all Thai. What if I was Chinese? Would I understand when they made some grammatical errors. In these cases I understood because I'm Thai." "So I didn"t always concentrate on their errors, just sometimes."

LU3: *"I didn"tconcentrate on any error, just kept listening."*

LU4: *"Always, whenever they made mistakes."*

LU5: *"Not so often. I looked at the errors in terms of understanding. For example, the third guy, the point of the unusual event was about the explosive devices but the guy couldn't get it at all when the interviewer asked what went wrong. This third guy didn"tsay anything about this. He talked about some other topics, even though this explosive device was the crucial part in this unusual event. The interviewer*

	was trying to link that there was another thing but he didn"t answer. So I had the feeling that it was an error, a misunderstanding." "I concentrated on the important parts and they missed them. They didn"t answer when they were questioned. These were errors." "I didn"t look at the grammar at all."

LU4 was the only linguistic/untrained rater who admitted that he *"always"* concentrated on the errors *"whenever they made mistakes"*. The other four raters (LU1, LU2, LU3 and LU5) said that they did not focus much on them. LU5 emphasized that he *"looked at the errors in terms of understanding"*. He *"didn"t bok at the grammar at all"*.

Table 4.103 shows the rating strategy of concentration on errors used by the operational/trained raters.

Table 4.103: The rating strategy of concentration on errors used by the operational/trained raters (OT)

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Much	Not much	Much	Not much	Much	Not much	Much	Not much	Much	Not much	
22. Concentration on errors		x		x		x	x		x		***OT1:*** *"I tried to minimize it but some mistakes were more difficult to get over." "I tried to write it down and then just forget about it and then try to find another part*

of negative. But there"s some situation where the negative stayed with me longer, then perhaps I noticed that, especially that part when he moved on to the next subject and I"m still remembering what he said on the last part."

OT2: *"I tried to pick up more to see if they"re consistent because we draw the border line between „thræ" and „four!"" "With the errors, if it"s more frequent or rarely. And if there were rarely, would they consistent on those kinds of errors like cluster sounds or ... the grammatical errors."*

OT3: *"For the first one I did the whole lot of that but afterwards I didn"t concentrate just on small errors. I just concentrated on the whole ... the whole English samples for number two and number three." "Because when number one I found that you"re concentrating so much on the errors, you didn"t really listen to the whole thing. And then what happened* (laughter)*...yeah ... that"s thereason for changing."*

	OT4: *"Very often because it was like my duty to do that."* ***OT5:*** *"How often? Almost all the time. Each time the speaker makes a mistake you"d note that down, remember that mistake and that would influence the rating that you would give. If he makes it fluently then it becomes his ... common mistake, right? So from ... instead of giving him a „four" then his grade may fall to a „thræ" when he makes too much mistake. So you would practically concentrate on all the mistakes he makes."*

OT1, OT2 and OT3 said that they did not concentrate much on the mistakes made by the candidates. OT1 *"tried to minimize it"* while OT2 *"tried to pick up more to see if they"re consistent"* and OT3 *"just concentrated on the whole"* because if *"you"re concentrating so much on the errors, you didn"treally listen to the whole thing".* OT4 and OT5 paid very much attention to the errors. OT5 also gave a very interesting remark that *"each time the speaker makes a mistake you"dnote that down, remember that mistake and that would influence the rating that you would give".*

Table 4.104 shows the rating strategy of concentration on errors used by the operational/untrained raters.

Table 4.104: The rating strategy of concentration on errors used by the operational/untrained raters (OU)

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Much	Not much	Much	Not much	Much	Not much	Much	Not much	Much	Not much	
22. Concentration on errors		x		x		x		x		x	***OU1:*** *"Initially I did but I didn"t afterward because they made little mistake and it was comprehensible."* ***OU2:*** *"No. I kept listening otherwise it"d be too tiring. When errors occur, I'd get them."* ***OU3:*** *"Not much. I didn"t concentrate on errors. I think I concentrated on the overall picture to see if they were intelligible."* ***OU4:*** *"The second guy seemed not to have enough attention." "His errors were obvious. That"s why I concentrated more on his errors." "The other made some errors too but I tried to understand that they were normal for Thais. I sometimes make some of those mistakes myself too."* ***OU5:*** *"I listened to what they said*

							and I knew if they made any mistake such as incorrect grammar or vocabulary. It wasn"t like concentration on any particular error. I just kept listening."

All operational/untrained raters seemed not to pay much attention to the errors made by the candidates. However, OU4 admitted that he concentrated more on the errors made by the second candidate because *"his errors were obvious"*. He said that the other candidates *made some errors too but"* he *"tried to understand that they were normal for Thais"*.

Table 4.105 shows the rating strategy used by the linguistic/trained raters in listening for types of errors.

Table 4.105: The rating strategy of used by the linguistic/trained raters (LT) in listening for types of errors

Raters	**Sub-themes**						**Meaning units**
	23. Types of errors that raters listened for (Number listed in priority)						
	Pronunciation	**Structure**	**Vocabulary**	**Fluency**	**Comprehension**	**Interactions**	
LT1	2	1	3	-	-	-	***LT1:*** *"Structure followed by pronunciation." "Also when he paused, was it because he didn"t know what to say or something else?" "Was it because he didn"t know the correct answer or he was stuck with the vocabulary?"*
LT2	1	1	1	2	2	2	***LT2:*** *"Mainly they were pronunciation, grammar and vocabulary. The other three*

							followed later."
LT3	1	1	1	1	2	2	***LT3:*** *"The first four criteria. They"reindicators of their language. The last two will follow later.*
LT4	-	1	2	-	-	-	***LT4:*** *"Mostly structure …" "I also listened partly to the vocabulary."*
LT5	1	1	1	2	2	2	***LT5:*** *"Pronunciation, structure and vocabulary because I feel that if his interaction is not good, it can be cured. If his comprehension is wrong, he can ask to repeat the question again which is not serious. Fluency is also not as serious as the first three criteria."*

The raters in the linguistic/trained group seemed to concentrate mainly on the first four criteria which were pronunciation, structure, vocabulary and fluency. For LT1 they were *"structure followed by pronunciation"*. *"Mainly they were pronunciation, grammar and vocabulary* for LT2. *"The first four criteria"* were focused by LT3 while LT4 concentrated *"mostly on structure"* and *"partly to the vocabulary"*. LT5 also pinpointed to *"pronunciation, structure and vocabulary because I feel if his interaction is not good, it can be cured. If his comprehension is wrong, he can ask to repeat the question again which is not serious. Fluency is also not as serious as the first three criteria."*

Table 4.106 shows the rating strategy used by the linguistic/untrained raters in listening for types of errors.

Table 4.106: The rating strategy of used by the linguistic/untrained raters (LU) in listening for types of errors

Raters	Sub-themes						Meaning units
	23. Types of errors that raters listened for (Number listed in priority)						
	Pronunciation	**Structure**	**Vocabulary**	**Fluency**	**Comprehension**	**Interactions**	
LU1	-	-	-	-	1	-	***LU1:*** *"I listened for comprehension."*
LU2	-	1	2	-	-	-	***LU2:*** *"Vocabulary and grammar." "More on grammar …"*
LU3	2	1	2	1	2	2	***LU3:*** *"I first looked at the fluency and structure." The other factors like vocabulary followed later."*
LU4	1	1	-	-	-	-	***LU4:*** *"Mostly they were pronunciation and grammatical structures …"*
LU5	2	-	-	2	1	2	***LU5:*** *"It was comprehension as the main point. I think what I could write the comments well were comprehension, fluency and interactions, and also pronunciation."*

Even though all of the linguistic/untrained raters are English teachers, not all of them concentrated on the errors made by the candidates in terms of grammar or structure. Just LU2, LU3 and LU4 focused mainly on grammatical structures. The other two (LU1

and LU5) said that they focused mainly on comprehension. It is worth noting that three raters (LU2, LU4 and LU5) in this batch admitted that they are *"not familiar with terms in this field"* (LU2). That was *"because vocabulary is field-specific so I wasn"tsure if they used them correctly or appropriately"* (LU4). LU5 even confessed that he *"didn"thave the ability to rate vocabulary because, as I told you, even though they paraphrased, I still didn"t know what"s the origin of their paraphrasing"*.

Table 4.107 shows the rating strategy used by the operational/trained raters in listening for types of errors.

Table 4.107: The rating strategy of used by the operational/trained raters (OT) in listening for types of errors

Raters	**Sub-themes**						**Meaning units**
	23. Types of errors that raters listened for (Number listed in priority)						
	Pronunciation	**Structure**	**Vocabulary**	**Fluency**	**Comprehension**	**Interactions**	
OT1	-	1	1	-	2	2	***OT1:*** *"Vocabulary and structure were the two that I notice on myself that I"mconcentrating. The least were comprehension and interactions"*
OT2	-	2	2	-	1	-	***OT2:*** *"Comprehension, I think I put it up in one of my top criteria …" "Grammar, I … probably because even though if you speak broken English but you can, maybe, paraphrase or you can … have a good vocabulary or you can use another word that, you know, just one word, you might change the whole thing.*

OT3	2	1	3	1	1	2	***OT3:*** *"Errors mostly on comprehension and the structure. Fluency as well. Interactions not so much …" "Oh! Well, pronunciation too but not that much." I didn"t mind so much about vocab …"*
OT4	1	2	3	-	4	4	***OT4:*** *"First I listened for pronunciation. Then I looked at the structure. The third was vocabulary. Comprehension and interactions come along with these."*
OT5	1	1	2	1	3	3	***OT5:*** *"Pronunciation, I think, is very important, right?" "I would concentrate on pronunciation and the fluency." "Um … structure first and then vocabulary because fluency and structure would go together …" "Comprehension and interactions would be the last things …"*

OT1 said that he concentrated more on *"vocabulary and structure"* while OT2 stated that the top criterion he focused on was comprehension. OT3 put his priority on comprehension, structure and fluency. OT4 had the same idea as OT5 that both of them said that they concentrated on pronunciation, structure and vocabulary.

Table 4.108 shows the rating strategy used by the operational/untrained raters in listening for types of errors.

Table 4.108: The rating strategy of used by the operational/untrained raters (OU) in listening for types of errors

Raters	Sub-themes 23. Types of errors that raters listened for (Number listed in priority)						Meaning units
	Pronunciation	**Structure**	**Vocabulary**	**Fluency**	**Comprehension**	**Interactions**	
OU1	1	1	1	-	-	-	***OU1:*** *"I concentrated on errors made in vocabulary, structure and pronunciation, not the other criteria."*
OU2	2	2	2	2	1	2	***OU2:*** *"Everything. Nothing in particular. It'd be serious if it was incomprehensible."*
OU3	-	-	-	-	1	1	***OU3:*** *"I think comprehension and interactions are the most two important things.*
OU4	2	1	2	2	2	2	***OU4:*** *"All of them. The most obvious was his sentence structure ..." "Okay, let"s put comprehension to number two." "It's not important to judge on the use of language fluency." " ... So I don"t mind this point, just comprehensible." "I just looked if I could generally understand them, if they could communicate under such circumstances."*

OU5	1	1	1	1	1	1	***OU5:*** *"I tried to award the scores according to ICAO guideline. I equally looked at all six criteria, not a special one."*

Two operational/untrained raters seemed to be cautious when they said that they *"tried to award the scores according to ICAO guideline"* (OU5) and focused on *"everything"* and *"nothing in particular"* (OU2). However, OU2 hinted that he might put more concentration on comprehension because *"it"d be serious if it was incomprehensible"*. OU1 typically concentrated on *"vocabulary, structure and pronunciation"* while OU3 thought that *"comprehension and interactions are the most two important things"*. Even though OU4 said in the beginning that he focused on *"all of them",* he admitted later that *"the most obvious was his sentence structure"* and he *"put comprehension to number two"*. However, OU4 accepted that he just looked if he *"could generally understand them, if they could communicate under such circumstances"* so he seemed to care most on the comprehension criterion.

Table 4.109 shows the rating strategy used by the linguistic/trained raters in considering the relatedness/relevance.

Table 4.109: The rating strategy of used by the linguistic/trained raters (LT) in considering the relatedness/relevance

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
24. Consideration on the relatedness/relevance of the content as a factor in their ratings	x		x		x		x		x		***LT1:*** *"Sometimes."* ***LT2:*** *"Partly if it concerned comprehension. If it"s not relevant, it means you don"t understand it." "Relevance and comprehension come together and also interactions. They come in as a package."*

	LT3: *"Yes." "Always."* ***LT4:*** *"It depends on the questions. It"s related to the content." "I didn"t concentrate on the content itself but the content leads to the language use."* ***LT5:*** *"Yes. Sometimes."*

Even though all linguistic/trained raters said that they considered the relatedness/relevance of the content as a factor in their ratings, only LT3 accepted that she *"always"* considered that. The others just considered the relatedness/relevance of the content as a sub-theme in their ratings for *"sometimes"* (LT1 and LT5) or *"partly"* (LT2). LT4"s remark was interesting that she considered it depending on the questions – if they were related to the content. She *"didn"tconcentrate on the content itself but the content leads to the language use."*

Table 4.110 shows the rating strategy used by the linguistic/untrained raters in considering the relatedness/relevance.

Table 4.110: The rating strategy of used by the linguistic/untrained raters (LU) in considering the relatedness/relevance

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
24. Consideration on the relatedness/relevance of the content as a factor in their ratings	x		x		x		x		x		***LU1:*** *"Yes. Very much. Very often."* ***LU2:*** *"Quite often. For example, when their answers were irrelevant to the questions." "Because I think this is serious in aviation."*

LU3: *"Yes but not so often. I focused on it in the interview part but not in the first part because I had no idea of that part. I didn"tknow if that was enough when one side spoke something and his counterpart answered. So I looked at the language only in this part. I turned to the content in the interview part."*

LU4: *"Yes, always." "If their answers were irrelevant to the questions, they affected their scores."*

LU5: *"Yes, I did. As I said, the third guy seemed to be good but he couldn"t get the point."*

Three linguistic/untrained raters confirmed that they *"always"* (LU4) or *"often"* (LU1 and LU2) considered the relatedness/relevance of the content as a sub-theme in their ratings. LU5 did not state clearly the degree of her consideration but she admitted that she did. LU3 confessed that she *"focused on it in the interview part but not in the first part"* because she *"had no idea of that part."* Contrary to LU3, LU2 who also did not have any background in aviation said that she considered the relevance of the content *"quite often"* because she thought that *"this is serious in aviation."*

Table 4.111 shows the rating strategy used by the operational/trained raters in considering the relatedness/relevance.

Table 4.111: The rating strategy of used by the operational/trained raters (OT) in considering the relatedness/relevance

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
24. Consideration on the relatedness/relevance of the content as a factor in their ratings		x	x		x		x		x		***OT1:*** *"No, I didn"t because in the training we received before they say that we are not here to grade on the specific procedure. We are here to assess their English proficiency so I didn"tuse that as negative or anything."* ***OT2:*** *"Relevance of the content? The one that I gave more consideration to ... the one that ... the urgency call like „mayday mayday mayday" or whatever because, I don"t know, when ... my point of view ... those are the ones where ... hey, you need to ... you sort of to be quick on that. Okay? Well, I was supposed to concentrate on the language but in some cases you need ... you know ... you need not just be able to use vocabulary." "Just in some particular cases like in an emergency situation." "I know I"m not supposed to but ..." (Laughter)* ***OT3:*** *"Like I said I listened to whole thing as a whole so the content was the most I pay attention on. So I say I paid pretty much attention on content." "Yes, always. That"s on comprehension."* ***OT4:*** *"Very little because everybody has his own*

	opinion." ***OT5:*** *"Well … it"s whether the answer is right or wrong. It"s … as raters, we are trained not to look at the answer. We are trained just to look if the answer matches the question being asked, right? And then look at the language being used, right? So if the answer is just along the line to the question then it"s fine. You don"t really want the correct answer. You want the correct language … for the answer." "I would if it"s a different kind of test, right? Not for the language assessment." "Very few. Very little."*

Most operational/trained raters seemed to understand the concept of language proficiency assessment thoroughly since they said something like *"in the training we received before they say that we are not here to grade on the specific procedure. We are here to assess their English proficiency …"* (OT1) and *"we are trained just to look if the answer matches the question being asked …" "And then look at the language being used …" "So if the answer is just along the line to the question then it"s fine. You don"t really want the correct answer. You want the correct language … for the answer."* (OT5). Nonetheless, OT2 admitted that he *"was supposed to concentrate on the language …" "… but in some cases you need … you know … you need not just be able to use vocabulary." "Just in some particular cases like in an emergency situation." "I know I'm not supposed to but …"* while OT3 *"listened to whole thing as a whole so the content was the most"* he paid attention on. He *"paid pretty much attention on content."* OT4 paid *"very little"* consideration but with another reason. It was not because of the concept of language proficiency assessment but *"because everybody has his own opinion."* That makes OT1 the only rater in this category who admitted that he did not take the relatedness or the relevance of the content into his consideration. OT5 who had the same idea of the language proficiency assessment as OT1 still accepted that he considered that in the degree of *"very few, very little."*

Table 4.112 shows the rating strategy used by the operational/untrained raters in considering the relatedness/relevance.

Table 4.112: The rating strategy of used by the operational/untrained raters (OU) in considering the relatedness/relevance

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
24. Consideration on the relatedness/relevance of the content as a factor in their ratings	x		x		x		x		x		***OU1:*** *"I did but they didn"t answer in that manner."* ***OU2:*** *"Yes. We can"t deny that our job is related to this."* ***OU3:*** *"Um ... yes." "I considered it but not so much. Just sometimes."* ***OU4:*** *"Always." "For example, the second guy could answer linguistically correct but it wasn"t the question he was asked." (Long pause) "I can"t think of the example."* ***OU5:*** *"Yes, I did. It concerned comprehension. If they ask you a question about one thing and you answer about some other thing, it means that you lack comprehension."*

All operational/untrained raters said that they considered the relatedness/relevance of the content as a sub-theme in their ratings. Three of them (OU1, OU2 and OU5) did not clearly state the degree of their considerations. OU3 *"considered it but not so much, just sometimes."* while OU4 *"always"* did it.

Table 4.113 shows the rating strategy used by the linguistic/trained raters in considering the quality of the content.

Table 4.113: The rating strategy of used by the linguistic/trained raters (LT) in considering the quality of the content

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
25. Consideration on the quality of the content as a factor in their ratings	x		x		x		x		x		***LT1:*** *"If his answer was irrelevant to the question, I would consider that in the area of comprehension."* ***LT2:*** *"I also consider it too but not much because some questions could be answered briefly, for example the questions in the last part."* ***LT3:*** *"Always."* ***LT4:*** *"Not quite."* ***LT5:*** *"Yes. Sometimes."*

All linguistic/trained raters said that they considered the quality of the content the candidates give as a sub-theme in their ratings with different degrees. LT1 would consider it in terms of comprehension if the candidate‟s answer *"was irrelevant to the question."* LT2, LT4 and LT5 considered it *"but not much" "not quite"* and *"sometimes"* respectively. Only LT3 stated that she *"always"* considered the quality of the content the candidates give as a factor in her rating.

Table 4.114 shows the rating strategy used by the linguistic/untrained raters in considering the quality of the content.

Table 4.114: The rating strategy of used by the linguistic/untrained raters (LU) in considering the quality of the content

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
25. Consideration on the quality of the content as a factor in their ratings	x		x		x		x		x		***LU1:*** *"Very often."* ***LU2:*** *"Yes because I have an opinion that, in aviation, if the answer has nothing to do with the question it may lead to an accident. So I think this issue is serious."* ***LU3:*** *"Yes, sometimes because of the same reasons as the previous question."* ***LU4:*** *"Yes, always."* ***LU5:*** *"I considered if the candidates answered according to the gist that the interviewer wanted to get." "I considered the relatedness more than the quality of the content." "It must be straight to the point. They must answer what they were asked." "It doesn"t have to be long or elaborate."*

All linguistic/untrained raters said that they considered the quality of the content the candidates give as a factor in their ratings with different degrees. It was *"very often"* and *"always"* for LU1 and LU4, just *"sometimes"* for LU3. LU2 sounded to take this more serious since she felt that *"in aviation, if the answer has nothing to do with the question it may lead to an accident."* LU5 *"considered if the candidates answered according to the gist that the interviewer wanted to get."* He also added that he *considered the relatedness*

more than the quality of the content." Moreover, *"it must be straight to the point"* and *"they must answer what they were asked." "It doesn"t have to be long or elaborate."*

Table 4.115 shows the rating strategy used by the operational/trained raters in considering the quality of the content.

Table 4.115: The rating strategy of used by the operational/trained raters (OT) in considering the quality of the content

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
25. Consideration on the quality of the content as a factor in their ratings		x	x		x		x		x		***OT1:*** *"No, not at all."* ***OT2:*** *"Yes, I do. Only in some cases, not often. I can say „rarely"."* ***OT3:*** *"Not really. I just wanna make sure that he can answer it."* ***OT4:*** *"Three out of five."* ***OT5:*** *"Very little."*

OT1 was still the only rater in the operational/trained category who said that he did not take the quality of the content as a factor in his rating. The others did it up to the different extents. It was *"rarely"* for OT2, *"not really"* for OT3, *"three out of five"* for OT4 and *"very little"* for OT5.

Table 4.116 shows the rating strategy used by the operational/untrained raters in considering the quality of the content.

Table 4.116: The rating strategy of used by the operational/untrained raters (OU) in considering the quality of the content

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
25. Consideration on the quality of the content as a factor in their ratings	x		x		x		x		x		***OU1:*** *"Yes, quite often."* ***OU2:*** *"Yes. If your answer is not according to the procedure, it may lead to something else as a consequence."* ***OU3:*** *"Yes. Sometimes."* ***OU4:*** *"Always." "The questions were not difficult. As a pilot, you should be able to answer them correctly."* ***OU5:*** *"Yes. This is the matter of getting „five" or „six". The guy who answers with more details, better quality would get higher score."*

All operational/untrained raters said that they considered the quality of the content the candidates give as a factor in their ratings with dissimilar strengths. It was *"quite often"* for OU1, *"sometimes"* for OU3 and *"always"* for OU4. OU2 and OU4 seemed to demonstrate his background as pilots by saying *"If your answer is not according to the procedure, it may lead to something else as a consequence"* and *"the questions were not difficult. As a pilot, you should be able to answer them correctly."* OU5 stated his

opinion which showed what would affect his ratings by saying *"this is the matter of getting „five" or „six". The guy who answers with more details, better quality would get higher score."*

Table 4.117 shows the rating strategy used by the linguistic/trained raters in considering the candidates" distinctive characteristics.

Table 4.117: The rating strategy of used by the linguistic/trained raters (LT) in considering the candidates" distinctive characteristics

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
26.Consideration on the candidates" distinctive characteristics	x			x	x			x	x		***LT1:*** *"If there was, it should have been the knowledge the candidates possessed in answering the questions. For example, candidate number one sounds „experience". He has experience in his job therefore he answered the questions instantly and I believed that those answers were correct. On the contrary, the other two candidates had no confidence in their answers."* ***LT2:*** *"No, not at all." "For example, when we talk about accent I'd care more about pronunciation. It is if his accent is strong enough to interfere with understanding." "Not because of his Thai accent."* ***LT3:*** *"Speech rate." "If they speak too fast, they'll miss a lot." "Accent doesn't matter." "There was one candidate, perhaps number two, but I'm not sure."*

	LT4: *"Um… no. But being whatever nationality has effects on everything in language usage especially pronunciation. I think more of it in this way. For example, being a Thai is like this. Talk like this. Translate word by word like this. I don"tknow what the Chinese do but these do not affect my rating."* ***LT5:*** *"No, not the accent." "I don"t want to use the term „bias". Let"s use the term „prɛference"." "Well, the term „prɛference" is not quite right." "I wouldn"t say they didn"t have any effect on my consideration. They did but very little."*

Linguistic/trained raters considered with different perspectives on the candidates" distinctive characteristics. LT1 said that she considered *"the knowledge the candidates possessed in answering the questions"* which made *"candidate number one sounds „ɛxperience"* while *"the other two candidates had no confidence in their answers."* LT3 said that she put the candidates" speech rate into her consideration. LT2 and LT4 stated that they did not consider any of the candidates" distinctive characteristics in their ratings. Even though LT5 did not clearly state the kind of characteristics, she accepted that those characteristics had some but very little effect on her consideration.

Table 4.118 shows the rating strategy used by the linguistic/untrained raters in considering the candidates" distinctive characteristics.

Table 4.118: The rating strategy of used by the linguistic/untrained raters (LU) in considering the candidates‟ distinctive characteristics

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
26. Consideration on the candidates‟ distinctive characteristics	x		x		x		x		x		***LU1:*** *“Yes. Their accent was very Thai.”* ***LU2:*** *“Being Thai.” “I usually give „thræ" for Thai people because I think what they do are good enough for Thai people.” “It doesn't mean that I give higher score because they are Thai.” “They would get higher score only when they do something better than that.”* ***LU3:*** *“Um … I looked at their responses. How fast they reacted.” “No, not in my rating.”* ***LU4:*** *“Yes. It was about the tone. I mean the confidence they had in answering the questions.” “As pilots, they should communicate clearly and confidently. They should have confidence in using language in their communication. I considered this in my rating too.”* ***LU5:*** *“They were secondary.” “I looked at their confidence as an important thing.” “I have special positive score for this confidence because if they are not confident in communicating, no matter how good they are, they wouldn't beable to convey what they want to.”*

All linguistic/untrained raters considered at least one distinctive characteristic of the candidates in their ratings. LU1 said that she considered the candidates‟ accent which *"was very Thai."* LT2 said that she considered another thing deeper than their accent. She considered the candidates as *"being Thai."* LU3 stated that she considered the candidates‟ responses but she refused that it had effect on her ratings. LU4 and LU5 looked at the same thing which was the candidates‟ confidence. LT4 even insisted that he considered this in his ratings while LT5 said that he had *"special positive score for this confidence."*

Table 4.119 shows the rating strategy used by the operational/trained raters in considering the candidates‟ distinctive characteristics.

Table 4.119: The rating strategy of used by the operational/trained raters (OT) in considering the candidates‟ distinctive characteristics

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
26. Consideration on the candidates‟ distinctive characteristics		x	x		x		x		x		***OT1:*** *"None of those. I tried to concentrate on what he was trying to say."* ***OT2:*** *"I'd concentrate on the accent that is difficult to understand, okay?" "I concentrate on the accent if it's comprehensible or not because you know that for an American or an Australian they probably don't have problem with grammatical structure, vocabulary. So ... I don't know ... I take the profile which nationality has a problem with."* ***OT3:*** *"I did consider ... flow ... the accent I didn't mind.*

	Fluency and ... yes ... fluency I think." "You mean did it make me bias in any way? Alright, initially may be." "Bias in terms of ... you don"t expect much if he doesn"t sound like an English or American or something. As soon as you listen ... okay, here we go ..." "Maybe accent." "Because he speaks with Thai accent and then you expect less. But in the end when you listen ... well, it"s not bad. Then you kind of bump the credit back again." "Because I know that he"s Thai, I start with this level in mind and then ... I start with ... what ... maybe level four and see if he"s up or down or see if he"s just on that." "So it"s accent and nationality."
	OT4: *"Probably their nationalities. Because I knew that they were Thai. When they spoke like that with such fluency, with such interactions, I considered that as „good enough"." "Because of their „Thainess". It wasn"t their native tongue." "Also their accents. Because they are Thai so I didn"t expect much from their accents. " "On the other hand, if they speak with very good accent, even though they are Thai, I"dconsider that too."*
	OT5: *"Being Thai, some people are very influenced by the Thai language so their pronunciation, their stress, their accent are very Thai which sometimes when you use it in English it"s very hard for me to understand. "Yes, of course. I"m influenced by the candidate"s accent, his pronunciation. All affect his score." "Yes. I consider accent as a part of pronunciation."*

OT1 was the only operational/trained rater who insisted that he did not consider any distinctive characteristics of the candidates in his ratings. The other raters accepted that they considered the accent of the candidates with different reasons. OT2 considered *"on the accent that is difficult to understand"*. OT3 said that he considered both the *"accent and nationality"* of the candidates. He accepted that he did not expect much if the candidate *"doesn"tsound like an English or American or something"* and *"Because I know that he"s Thai, I start with this level in mind and then ... I start with ... what ... maybe level four and see if he"s up or down or see if he"sjust on that."* OT4 also said that he considered the candidates" nationalities and accents. He even admitted that he *"didn"t expect much from their accents"* because they were Thai. OT5 acknowledged that he was *"influenced by the candidate"s accent, his pronunciation"* and *"all affect his score."*

Table 4.120 shows the rating strategy used by the operational/untrained raters in considering the candidates" distinctive characteristics.

Table 4.120: The rating strategy of used by the operational/untrained raters (OT) in considering the candidates" distinctive characteristics

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
26. Consideration on the candidates" distinctive characteristics	x		x		x		x			x	***OU1:*** *"I was negatively marked by the third candidate"s pronunciation. If I were not Thai, I wouldn"t be able to get what he said." "His accent was so Thai." "It was a negative impression that he was too Thai." "Furthermore, his voice was not clear, it wasn"t sharp."* ***OU2:*** *"Partly, but not much." "We"re not native speakers so I don"tcare about the accent." "Because we*

	are Thai so I didn't expect much." ***OU3:*** *"If I have to choose, it should be voice and tone." "I think if somebody has bad grammar but it'd be fine if he is calm, speaks slowly and clearly. I think this is important."* ***OU4:*** *"Yes. Being Thai. Because of that I have excuses for every mistake in any criterion they made. I'd like to add that because they are Thai like me."* ***OU5:*** *"No, I didn't. For example, all of the candidates spoke with Thai accent but it didn't matter." "So I didn't consider such characteristics."*

The candidates' accent was the distinctive characteristic which was accepted by OU1 that he had a *"negative impression"* on the third candidate that *"he was too Thai."* OU2 had an opposite idea about the candidates' Thai accent. He said *"because we are Thai so I didn't expect much."* OU4 had the same perspective on the candidates' Thai accent. He had *"excuses for every mistake in any criterion they made" "because they are Thai like me."* OU3 said that he considered *"voice and tone"* of the candidates. He thought that *"if somebody has bad grammar but it'd be fine if he is calm, speaks slowly and clearly".* OU5 was the only rater in this category who *"didn't consider such characteristics."*

Table 4.121 shows the rating strategy used by the linguistic/trained raters in putting equal weight on all six criteria.

Table 4.121: The rating strategy of used by the linguistic/trained raters (LT) in putting equal weight on all six criteria

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
27. Putting equal weight on all six criteria	x		x		x			x		x	***LT1:*** *(After a period of pondering) "I never think of that but I think it should be equally."* ***LT2:*** *"Yes."* ***LT3:*** *"Equally." "I gave importance to the six criteria equally but while listening, because the first four criteria were what they produced then they"d show the meanings and how they interacted with the problems."* ***LT4:*** *"I think I weighed more on the first three criteria because they lead to the latter three."* ***LT5:*** *"No. I weighed more on the first three." "They are important but it doesn"t mean that if these three are good, I would not consider the other three. They are a little more weighted but not that significant. It means if the last three are not good and the first three are good, it would affect the final score. No, it wouldn"t." "They are not totally weighted more than the last three."*

LT1, LT2 and LT3 said that they weighted all six criteria equally in their ratings. LT4 and LT5 accepted that they weighted more on *"the first three criteria" "because they lead to the latter three."* However, LT5 added that *"they are a little more weighted but not that significant."*

Table 4.122 shows the rating strategy used by the linguistic/untrained raters in putting equal weight on all six criteria.

Table 4.122: The rating strategy of used by the linguistic/untrained raters (LU) in putting equal weight on all six criteria

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
27. Putting equal weight on all six criteria		x	x		x		x			x	***LU1:*** *"I weighed structure and vocabulary less than others ..." "I can"t saybut, let"s say, I didn"t concentrate much on structure and vocabulary." "I checked the comprehension when they answered, to see how much they understood."* ***LU2:*** *"My overall score is the majority of six criteria." "For example, this candidate gets „four"for overall because he gets „four"in most criteria." "But I think I weighed each criterion equally."* ***LU3:*** *"Yes."* ***LU4:*** *"Yes."* ***LU5:*** *"I put more weight on comprehension. So I didn"tweigh it equally." "As I said, it was because of my background. It"s impossible if I want to weigh on vocabulary because I didn"t know much vocab. For example, the word „taxi", I didn"t know what it was." "I have limited background in aviation." "So I*

	rated vocabulary in terms of general vocabulary instead of those specific technical terms."

LU2, LU3 and LU4 said that they weighted each criterion equally. LU1 said that she *"weighted structure and vocabulary less than others"* while LU5 *"put more weight on comprehension"*. LU5 added that she had *"limited background in aviation"*. That was the reason why she did not put much weight on vocabulary which was quite specific in terms of technical terms.

Table 4.123 shows the rating strategy used by the operational/trained raters in putting equal weight on all six criteria.

Table 4.123: The rating strategy of used by the operational/trained raters (OT) in putting equal weight on all six criteria

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
27. Putting equal weight on all six criteria		x	x			x		x	x		***OT1:*** *"As I said, the comprehension and interactions I considered those less …"* ***OT2:*** *"So, yes, I think I weighed each criterion equally."* ***OT3:*** *"Ah … I know that I looked at the … structure … comprehension … and pronunciation and … fluency. One, two, three, four … these four looked at the whole lot more than interactions and vocab."* ***OT4:*** *"No, not equally." "For final scores, I weighed more on comprehension and structure."* ***OT5:*** *"Yes."*

Two raters (OT2 and OT5) in the operational/trained group said that they weighed all six criteria equally. OT1 stated that he considered comprehension and interactions less than the other four criteria. OT3 admitted that he put more weight on pronunciation, structure, fluency and comprehension while OT4 accepted that he did not put them equally. He *"weighed more on comprehension and structure."*

Table 4.124 shows the rating strategy used by the operational/untrained raters in putting equal weight on all six criteria.

Table 4.124: The rating strategy of used by the operational/untrained raters (OU) in putting equal weight on all six criteria

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
27. Putting equal weight on all six criteria		x	x			x		x	x		***OU1:*** *"Yes." "Well…I might put more weight on comprehension and interactions because I give weight on comprehensibility, on communicability." "I think they are more crucial factors in communication."* ***OU2:*** *"I think I did equally."* ***OU3:*** *"No, not equally. They are not much different but I think I put most weight on comprehension for the same reason I mentioned. If you don't understand, what's the point? Interactions come next. What are interactions? If you don"t comprehend, you"ve got to interact. Right? You have to initiate, to ask." "I think I don"tmind much about structure. I put it last on the list. The others like vocabulary and pronunciation are auxiliary factors. I may be wrong. I don't know."* ***OU4:*** *"No, not equally. As I said, I weighed some criteria less*

	than others." "Because I don"t think they are all equally important, not in terms of comprehensibility." ***OU5:*** *"Yes because they are equally important. If you"regood in one aspect but bad in others, it"suseless to communicate with others. They must come together."*

OU2 and OU5 were two raters in the operational/untrained batch who said that they weighted all six criteria equally. OU1 *"put more weight on comprehension and interactions"* while OU3 *"put most weight on comprehension"*. OU4 did not clearly state what criteria he weighted more but he insisted that he did not put them equally because he did not *"think they are all equally important, not in terms of comprehensibility."*

Table 4.125 shows the degrees of the test task difficulty as considered by the linguistic/trained raters.

Table 4.125: The degrees of the test task difficulty as considered by the linguistic/trained raters (LT)

Sub-themes	**Rater LT1**			**Rater LT2**			**Rater LT3**			**Rater LT4**			**Rater LT5**			**Meaning units**
	E	**D**	**N**	**E**	**D**	**N**	**E**	**D**	**N**	**E**	**D**	**N**	**E**	**D**	**N**	
28. Degrees of test tasks i.e. E = Easy D = Difficult N = Neither			x			x			x			x			x	***LT1:*** *"Not difficult, not easy because it was a combination of ease and difficulty."* ***LT2:*** *"Not easy, not difficult."* ***LT3:*** *"Not easy, not difficult." "Quite appropriate because the tasks concerned their jobs directly." "The questions were neither too easy nor too difficult."*

	LT4: *"Not too easy, not too difficult because no matter what level the test-takers are, they''ll reœive the same questions."* ***LT5:*** *"Not difficult, not easy because it was appropriate to the situation."*

Even though they had different reasons, none of the linguistic/trained rater thought the test tasks were too easy or too difficult. LT1 thought, *"it was a combination of ease and difficulty"*. LT2 did not specify any particular reason. LT3 thought, *"the tasks concerned their jobs directly."* LT4 thought, *"no matter what level the test-takers are, they''llreceive the same questions"* while LT5 said, *"it was appropriate to the situation."*

Table 4.126 shows the degrees of the test task difficulty as considered by the linguistic/untrained raters.

Table 4.126: The degrees of the test task difficulty as considered by the linguistic/untrained raters (LU)

Sub-themes	Rater LU1			Rater LU2			Rater LU3			Rater LU4			Rater LU5			Meaning units
	E	**D**	**N**	**E**	**D**	**N**	**E**	**D**	**N**	**E**	**D**	**N**	**E**	**D**	**N**	
28. Degrees of test tasks i.e. E = Easy			x			x		x				x		x		***LU1:*** *"Not easy." (Hesitant) "Um…intermediate because they were patterns and codes which are usually used by pilots then it''s easy, but it was difficult when unusual situations happened."*
D =																***LU2:*** *"Not difficult, not easy because the real situations they*

Difficult N = Neither	*normally experience should be the same."* ***LU3:*** *"I think it was difficult because they had to respond instantly, especially the first part, not the interview part."* ***LU4:*** *"Not difficult because most of the language they used were phraseologies which pilots knew how to respond and most of them were repetition. They repeated the orders between the air traffic controller and the pilot. The language in this part wasn"t difficult." "The interview part wasn"t difficult because the interviewer spoke slowly." "Not easy because there may be some factors in some situations which may influence the way pilots decide what to communicate. Pilots have to solve the problems and at the same time they have to think of the words to communicate with the air traffic controllers. So the overall is not difficult and not easy. It requires quite an interaction."* ***LU5:*** *"Quite difficult, perhaps because I'm not in this field. So I feel it"srather difficult."*

Three linguistic/untrained raters said that the test tasks were neither too easy nor too difficult because *"they were patterns and codes which are usually used by pilots then it"s easy but it was difficult when unusual situations happened"* (LU1), *"the real situations they normally experience should be the same"* (LU2) and *"most of the language they used were phraseologies which pilots knew how to respond and most of them were repetition"* (LU4). However, the other two thought differently. Those were because *"they had to respond instantly"* (LU3) and *"I'm not in this field"* (LU5).

Table 4.127 shows the degrees of the test task difficulty as considered by the operational/trained raters.

Table 4.127: The degrees of the test task difficulty as considered by the operational/trained raters (OT)

Sub-themes	Rater OT1			Rater OT2			Rater OT3			Rater OT4			Rater OT5			Meaning units
	E	D	N	E	D	N	E	D	N	E	D	N	E	D	N	
28. Degrees of test tasks i.e. E = Easy D = Difficult N = Neither			x	x			x				x		x			***OT1:*** *"I thought it was fair. I thought it wasn"ttoo hard, it wasn"t too difficult ..." "I think it"s difficult when the subject has to come up with something that they might not be too familiar with."* ***OT2:*** *"I think they were easy in all three parts." "Because I think I have broader vocabulary and I could answer the tasks being confronted."* ***OT3:*** *"That"s alright. It wasn"t difficult. It was an easy too, you know. I think it was just right. It"s okay. It"s good."* ***OT4:*** *"Quite difficult. I give eight out of ten because some tasks are not everyday life topics." "... It"s even more difficult to say it in a foreign language."* ***OT5:*** *"Oh! Okay. If that"s like that, it"s easy." "Because taking two courses of rating already ..." "For me it"dquite easy, but for them it depends. ...." "I can"tsay if it"s too easy or too hard. It depends on the test-takers."*

Three operational/trained raters thought that the tasks were easy because *"I have broader vocabulary and I could answer the tasks being confronted." "If I am an interviewee, I can answer all questions easily."* (OT2). OT3 did not present any specific reason for his opinion. OT5 said that the tasks were easy for him but *"it depends on the test-takers."* OT4 was the only rater who regarded

the test tasks as difficult because *"some tasks are not everyday life topics."* OT1 thought that *"it was fair"* but *"it"s difficult when the subject has to come up with something that they might not be too familiar with."*

Table 4.128 shows the degrees of the test task difficulty as considered by the operational/untrained raters.

Table 4.128: The degrees of the test task difficulty as considered by the operational/untrained raters (OU)

Sub-themes	Rater OU1			Rater OU2			Rater OU3			Rater OU4			Rater OU5			Meaning units
	E	D	N	E	D	N	E	D	N	E	D	N	E	D	N	
28. Degrees of test tasks i.e. E = Easy D = Difficult N = Neither			x			x		x				x			x	***OU1:*** *"Not difficult, not easy because they directly concern with their job. They are their routine duties." "It may be difficult when they had to explain something, bomb, something like that."* ***OU2:*** *"Not easy, not difficult."* ***OU3:*** *"I think it was difficult for test-takers. It"s not too hard for me." "That"s because they"re all open-ended questions which we, pilots, are not comfortable with them. We are unfamiliar to them."* ***OU4:*** *"Moderate, not difficult, not easy." "Because there were various kinds of questions. The questions were mixed."* ***OU5:*** *"For me, it was neither difficult nor easy because they assessed in many aspects. Okay it"snot too easy because it"s not our language but it"s not too difficult because it was aviation related. They should know about it." "It may not be*

	easy because we don"tuse it everyday. We use it only during flight and what we use are standardized words and phrases. If we work in a multi-national company and we have to use English everyday, it may not be so hard."

Almost all operational/untrained raters thought that the test tasks were neither easy nor difficult. However, *"it may be difficult when they had to explain something, bomb, something like that"* (OU1). OU3 regarded them as *"it was difficult for test-takers"* but *"it"snot too hard."* For him while OU5 said that *"it was neither difficult nor easy because they assessed in many aspects"* and *"it"s not too easy because it"snot our language but it"s not too difficult because it was aviation related."* For OU4 it was *"moderate, not difficult, not easy" "because there were various kinds of questions"* and *"the questions were mixed."* OU2 did not specify his reasons.

Table 4.129 shows the duration of the speech samples as considered by the linguistic/trained raters.

Table 4.129: The duration of the speech samples as considered by the linguistic/trained raters (LT)

Sub-themes	Rater LT1			Rater LT2			Rater LT3			Rater LT4			Rater LT5			Meaning units
	S	L	A	S	L	A	S	L	A	S	L	A	S	L	A	
29. Duration of the speech samples i.e. S = Too short L = Too long A = Appropriate		x				x			x			x			x	***LT1:*** *"Overall, it was a little too long."* ***LT2:*** *"Not too long, not too short."* ***LT3:*** *"Not too long. It was appropriate but I think the computer mediated part was a little too long. The interview part should be extended."* ***LT4:*** *"Its duration is not too long but the parts that*

	we use to rate, to judge the candidates are …okay." ***LT5:*** *"They were okay."*

Four out of five linguistic/trained raters (LT2, LT3, LT4 and LT5) said that the duration of the speech samples were *"okay"*. However, LT3 commented that *"the computer mediated part was a little too long"* and *"the interview part should be extended"*. LT1 was the only rater who said that *"overall, it was a little too long."*

Table 4.130 shows the duration of the speech samples as considered by the linguistic/untrained raters.

Table 4.130: The duration of the speech samples as considered by the linguistic/untrained raters (LU)

Sub-themes	Rater LU1			Rater LU2			Rater LU3			Rater LU4			Rater LU5			Meaning units
	S	L	A	S	L	A	S	L	A	S	L	A	S	L	A	
29. Duration of the speech samples i.e. S = Too short L = Too long A = Appropriate			x			x		x				x			x	***LU1:*** *"Not too long, not too short. Appropriate."* ***LU2:*** *"Not too short, not too long." "It depends on how fluent they answer the questions …"* ***LU3:*** *"It wasn't short, rather long."* ***LU4:*** *"Not too short, not too long." "They were appropriate for each part."* ***LU5:*** *"The duration of each sample was not equal. The first was too long. The third was a little too short. The second one seemed to be alright."*

Three linguistic/untrained raters (LU1, LU2 and LU4) stated that the speech sample duration was *"not too long, not too short"*. LU2 added that *"it depends on how fluent they answer the questions"* and *"they didn"tanswer long but it took a long time for them to answer."* LU3 had a conflicting idea that *"it wasn"tshort, rather long."* LU5 had the differing thought for each sample that *"the duration of each sample was not equal. The first was too long. The third was a little too short. The second one seemed to be alright."* It could be concluded that the overall length was *"alright."*

Table 4.131 shows the duration of the speech samples as considered by the operational/trained raters.

Table 4.131: The duration of the speech samples as considered by the operational/trained raters (OT)

Sub-themes	Rater OT1			Rater OT2			Rater OT3			Rater OT4			Rater OT5			Meaning units
	S	L	A	S	L	A	S	L	A	S	L	A	S	L	A	
29. Duration of the speech samples i.e. S = Too short L = Too long A = Appropriate			x			x			x			x			x	***OT1:*** *"The actual samples were not that long but they had a lot of gaps …" "If you cut that out, I think it would be fine."* ***OT2:*** *"I think they were good length."* ***OT3:*** *"I think the speech samples depend on how good that person"s English was." "I thought it was appropriate."* ***OT4:*** *"The role play part was too long while the interview part was too short." "It"s okay in average."* ***OT5:*** *"Um …at the end I think it was fine." "So I think maybe a bit too short."*

Four operational/trained raters agreed that the duration of the speech samples was *"fine"* or *"okay" "in average"*. OT1 also pointed out that *"the actual samples were not that long but they had a lot of gaps especially in the first one."* OT3 added that *"the speech samples depend on how good that person's English was"*. OT5 changed his mind from *"at the end I think it was fine"* to *"maybe a bit too short."* That made him the only rater in this group who thought that the speech samples duration was too short.

Table 4.132 shows the duration of the speech samples as considered by the operational/untrained raters.

Table 4.132: The duration of the speech samples as considered by the operational/untrained raters (OU)

Sub-themes	Rater OU1			Rater OU2			Rater OU3			Rater OU4			Rater OU5			Meaning units
	S	L	A	S	L	A	S	L	A	S	L	A	S	L	A	
29. Duration of the speech samples i.e. S = Too short L = Too long A = Appropriate		x			x			x			x			x		***OU1:*** *"It was rather long but it had to be long to judge them."* ***OU2:*** *"The first was too long."* ***OU3:*** *"I think if the radiotelephony part is shorter, it"ll be okay."* ***OU4:*** *"I think part one was too long." "The other parts are okay. Part one should be shorter and give more time to part two because the rating is based on part two."* ***OU5:*** *"The first one was too long. The latter two were okay."*

Four operational/untrained raters (OU2, OU3, OU4 and OU5) were unanimous that *"the first one was too long"*. OU1 was the only rater in this category who did not clearly stated about that first part. He just merely said that *"it was rather long"* but he agreed that *"it had to be long to judge them."* This might be because they are all pilots who are very familiar with the radiotelephony. So they saw this part as a waste of time. It also made them think that the overall duration was too long.

Table 4.133 shows the appropriate duration of the speech samples as considered by the linguistic/trained raters.

Table 4.133: The appropriate duration of the speech samples as considered by the linguistic/trained raters (LT)

Sub-themes	Rater LT1	Rater LT2	Rater LT3	Rater LT4	Rater LT5	Meaning units
30. Appropriate duration of the speech samples (minutes)	20-25	25-30	30-40	35-40	20	***LT1:*** *"Twenty to twenty five minutes. Maximum is thirty minutes."* ***LT2:*** *"It should be around 25 to 30 minutes depending on the questions that we ask. How deep we want to probe? For example, if we ask just very plain questions throughout 25 minutes, it wouldn"t beratable."* ***LT3:*** *"Around thirty to forty minutes."* ***LT4:*** *"Thirty-five to forty minutes including the radiotelephony part."* ***LT5:*** *"It should be the same length as the third one which was around twenty minutes."*

The appropriate duration of the speech samples varied among five linguistic/trained raters. The minimum was 20 minutes LT1 and LT5) and the maximum was 40 minutes (LT3 and LT4).

Table 4.134 shows the appropriate duration of the speech samples as considered by the linguistic/untrained raters.

Table 4.134: The appropriate duration of the speech samples as considered by the linguistic/untrained raters (LU)

Sub-themes	Rater LU1	Rater LU2	Rater LU3	Rater LU4	Rater LU5	Meaning units
30. Appropriate duration of the speech samples (minutes)	30	15-30	15-20	20-30	30	***LU1:*** *"What do you want to use it for?" "For proficiency test, it should be approximately half an hour."* ***LU2:*** *"Not more than 15 minutes." "It also depends on the task type." "These 15 minutes must cover the TLU (Target Language Use) that pilots really use." "It shouldn"t belonger than half an hour."* ***LU3:*** *"Around fifteen to twenty minutes." "It depends on how well you can make them speak. It wouldn"t be enough if the test-taker answers just one or two words. So it depends on the interviewer to encourage them to speak."* ***LU4:*** *"The third was quite alright. It"s around twenty to thirty minutes."* ***LU5:*** *"Around thirty minutes."*

The appropriate duration of the speech samples also varied among five linguistic/untrained raters. The minimum was 15 minutes (LU2 and LU3) and the maximum was 30 minutes (LU1, LU2, LU4 and LU5).

Table 4.135 shows the appropriate duration of the speech samples as considered by the operational/trained raters.

Table 4.135: The appropriate duration of the speech samples as considered by the operational/trained raters (OT)

Sub-themes	Rater OT1	Rater OT2	Rater OT3	Rater OT4	Rater OT5	Meaning units
30. Appropriate duration	30-45	30-40	20-30	30	30-40	***OT1:*** *"About 30 to 45 minutes."*

of the speech samples (minutes)	***OT2:*** *"I think these were roughly around forty minutes." "I think it shouldn"t bemore than that. I think thirty minutes is sufficient."* ***OT3:*** *"Yes, twenty to thirty minutes I think it"s enough."* ***OT4:*** *"Thirty minutes with full content. I mean not full of those gap fillers like „well", „er", „ah"."* ***OT5:*** *"Thirty to forty minutes is fine."*

The appropriate duration of the speech samples varied among five operational/trained raters. The minimum was 20 minutes (OT3) and the maximum was 45 minutes (OT1).

Table 4.136 shows the appropriate duration of the speech samples as considered by the operational/untrained raters.

Table 4.136: The appropriate duration of the speech samples as considered by the operational/untrained raters (OU)

Sub-themes	Rater OU1	Rater OU2	Rater OU3	Rater OU4	Rater OU5	Meaning units
30. Appropriate duration of the speech samples (minutes)	20-30	30	30	25	20	***OU1:*** *"Twenty to thirty minutes."* ***OU2:*** *"It should be around half an hour."* ***OU3:*** *"I think around thirty minutes is okay."* ***OU4:*** *"Totally not more than twenty-five minutes."* ***OU5:*** *"Around twenty minutes should be*

	enough."

The appropriate duration of the speech samples varied among five operational/untrained raters. The minimum was 20 minutes (OU1 and OU5) and the maximum was 30 minutes (OU2 and OU3).

Table 4.137 shows the linguistic/trained raters" (LT) opinions if rating three speech samples was too much.

Table 4.137: The linguistic/trained raters" (LT) opinions if rating three speech samples was too much

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
31. Rating three speech samples consecutively was too much	x			x		x		x	x		***LT1:*** *"It shouldn"tbe more than two."* ***LT2:*** *"It was okay."* ***LT3:*** *"It would be okay if I took a short brake."* ***LT4:*** *"No."* ***LT5:*** *"Yes, too much for me."*

Three linguistic/trained raters (LT2, LT3 and LT4) thought that rating three speech samples consecutively was *"okay"* while the other two thought that *"it shouldn"t bemore than two"* (LT1) or *"too much"* (LT5).

Table 4.138 shows the linguistic/untrained raters‟ (LU) opinions if rating three speech samples was too much.

Table 4.138: The linguistic/untrained raters‟ (LU) opinions if rating three speech samples was too much

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
31. Rating three speech samples consecutively was too much		x		x	x			x	x		***LU1:*** *"Appropriate."* ***LU2:*** *"No."* ***LU3:*** *"Quite too much."* ***LU4:*** *"No, not for me."* ***LU5:*** *"A little."*

Three linguistic/untrained raters (LU1, LU2 and LU4) thought that rating three speech samples consecutively was not too much while the other two thought that it was *"quite too much"* (LU3) or *"a little"* too much (LU5).

Table 4.139 shows the operational/trained raters‟ (OT) opinions if rating three speech samples was too much.

Table 4.139: The operational/trained raters‟ (OT) opinions if rating three speech samples was too much

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
31. Rating three speech samples consecutively was too much	x		x		x		x			x	***OT1:*** *"It depends really on how you rate each one." "I think three would have been too much. Two would have been good."* ***OT2:*** *"Yes."* ***OT3:*** *"Yes. I think two at any one time and then break. It"s much better."* ***OT4:*** *"Too much."* ***OT5:*** *"Was too much? Um ... I guess not but if it"s a full day to do this, right? I could have done one and then ..."*

Almost all operational/trained raters (OT1, OT2, OT3 and OT4) thought that rating three speech samples consecutively was too much. (*"Two would have been good"* – OT1, *"Yes"* – OT2 and OT3, *"Too much"* – OT4) Only OT5 did not think so.

Table 4.140 shows the operational/untrained raters‟ (OU) opinions if rating three speech samples was too much.

Table 4.140: The operational/untrained raters‟ (OU) opinions if rating three speech samples was too much

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
31. Rating three speech samples consecutively was too much	x		x			x	x			x	***OU1:*** *"Two should be enough but this was my first time. After this, I think three are not too many."* ***OU2:*** *"Yes. I got tired when listening to the third candidate."* ***OU3:*** *"I think it was nice. I can say that the first guy might not be rated accurately. Comparison is human nature. If you say you don"t compare, you definitely lie. I got clearer picture when I rated the second and the third guy."* ***OU4:*** *"If I have to write the comments like this, it was too much."* ***OU5:*** *"If there"s no more in the same day, it"s okay."*

Three operational/untrained raters thought that rating three speech samples consecutively was too much. OU1 said that *"two should be enough*". OU2 complained that he *"got tired when listening to the third candidate"* and OU4 stated that *"it was too much"* if he had to *"write the comments like this".* OU5 did not have any complaint *"if there"sno more in the same day".* OU3 was the only one who thought *"it was nice"* because he *"got clearer picture when* he *rated the second and the third guy".*

Table 4.141 shows the linguistic/trained raters‟ (LT) opinions concerning the maximum number of the speech samples that should be rated in one day.

Table 4.141: The linguistic/trained raters‟ (LT) opinions concerning the maximum number of the speech samples that should be rated in one day

Sub-themes	Rater LT1	Rater LT2	Rater LT3	Rater LT4	Rater LT5	Meaning units
32. The maximum number of the speech samples that should be rated in one day	5	6	4-6	6	4	***LT1:*** *"Maximum is five per day." "It also depends on the speech samples if they are easy or difficult to rate, if they are complicated or not."* ***LT2:*** *"I think it shouldn"t be more than six. It"s sort of three in the morning and three in the afternoon." "Otherwise raters may be too tired and the results may be unreliable."* ***LT3:*** *"It depends on how difficult to rate. If it"s difficult, four is enough." "So it depends on the candidates" proficiency, if they are good, five or six wouldn"t be a problem. But if they are weak, four is tough enough."* ***LT4:*** *"If they are easy like these, six are fine. Three in the morning and three in the afternoon."* ***LT5:*** *"Two in the morning and two in the afternoon."*

The linguistic/trained raters did not agree on the maximum number of the speech samples that should be rated in one day. LT1 said that *"maximum is five per day"*. It *"shouldn"t bemore than six"* for LT2. LT3 had the opinion that *"it depends on how difficult to rate"*. She explained that *"if they are good, five or six wouldn"tbe a problem. But if they are weak, four is tough enough"*. *"Six are fine"* for LT4 but only four which is *"two in the morning and two in the afternoon"* for LT5.

Table 4.142 shows the linguistic/untrained raters‟ (LU) opinions concerning the maximum number of the speech samples that should be rated in one day.

Table 4.142: The linguistic/untrained raters‟ (LU) opinions concerning the maximum number of the speech samples that should be rated in one day

Sub-themes	Rater LU1	Rater LU2	Rater LU3	Rater LU4	Rater LU5	Meaning units
32. The maximum number of the speech samples that should be rated in one day	6	6	1-2	6	6	***LU1:*** *"Three in the morning and three in the afternoon. Five would be too much."* ***LU2:*** *"Three in the morning and other three in the afternoon are okay - for rating only."* ***LU3:*** *"One or two in a day, if it"sthis long." "I felt that three hours was short so I think one day should be alright."* ***LU4:*** *"I think two in the morning and two in the afternoon are appropriate." "Three in the morning and three in the afternoon is the maximum limit."* ***LU5:*** *"It should be thirty minutes for one candidate, then a ten-minute break. The best practice should be three in the morning and three in the afternoon."*

Almost all linguistic/untrained raters (LU1, LU2, LU4 and LU5) consented that the maximum number of the speech samples that should be rated in one day was six. Only LU3 thought that it should be just *"one or two in a day, if it"s thislong"*.

Table 4.143 shows the operational/trained raters" (OT) opinions concerning the maximum number of the speech samples that should be rated in one day.

Table 4.143: The operational/trained raters" (OT) opinions concerning the maximum number of the speech samples that should be rated in one day

Sub-themes	Rater OT1	Rater OT2	Rater OT3	Rater OT4	Rater OT5	Meaning units
32. The maximum number of the speech samples that should be rated in one day	-	4	6	4	4	***OT1:*** *"It can"tbe said how many should be done in a day. It depends on the level as I said."* ***OT2:*** *"Two would be alright." "Two at a time. In a day, two by two."* ***OT3:*** *"Yes. You can do six in one day because after two you can break. If you did three in the morning, you can do one, break, one, break, one, break." "Three in the morning, three in the afternoon in one day."* ***OT4:*** *"Two. I mean one then a brake, another one and another brake. If we do it consecutively, there may be a comparison because we just finish the first one. Even though we know that shouldn't be done but it can"t help because it"s justdone." "Four ratings are the maximum in a day."* ***OT5:*** *"In one day I guess it"sabout three to four. We did three in half a day, right? Because of the time limit. But if I was given full day, I'd do two in the morning and two in the afternoon."*

Three operational/trained raters (OT2, OT4 and OT5) were consistent that rating four speech samples in a day was the maximum. OT3 thought he could handle six samples in a day while OT1 did not state clearly how many it should have been. He just mentioned, *"it depends on the level"*.

Table 4.144 shows the operational/untrained raters‟ (OU) opinions concerning the maximum number of the speech samples that should be rated in one day.

Table 4.144: The operational/untrained raters‟ (OU) opinions concerning the maximum number of the speech samples that should be rated in one day

Sub-themes	Rater OU1	Rater OU2	Rater OU3	Rater OU4	Rater OU5	Meaning units
32. The maximum number of the speech samples that should be rated in one day	6	4	4	3	6	***OU1:*** *"Three in the morning and three in the afternoon."* ***OU2:*** *"Two in the morning and other two in the afternoon."* ***OU3:*** *"It's the diminishing returns." "Two in the morning and two in the afternoon should be better."* ***OU4:*** *"Let"s say two in the morning and one in the afternoon because people normally get tired in the afternoon. There"ll be some other factors which affect in the afternoon. It's more energetic in the morning."* ***OU5:*** *"In real assessment three in the morning and three in the afternoon should be fine."*

Two operational/untrained raters (OU1 and OU5) thought that the maximum number of the speech samples that should be rated in one day was six while the other two (OU2 and OU3) thought it was four. OU4 was the only one who thought of a number less than that. In his opinion, *"people normally get tired in the afternoon"* and *"it"s more energetic in the morning"*. Moreover, he supposed that *"there"ll be some other factors which affect in the afternoon"*. That is why he proposed to rate simply three speech samples in a day.

Table 4.145 shows the linguistic/trained raters‟ (LT) considerations concerning the interviewers‟ accommodation.

Table 4.145: The linguistic/trained raters‟ (LT) considerations concerning the interviewers‟ accommodation

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
33. The interviewers/interlocutors tried to help/accommodate the candidate during the test		x	x		x		x		x		***LT1:*** *"No. He knew when to repeat the questions again."* ***LT2:*** *"Sometimes because if the interviewee seemed not to understand, he would tailor the language he used or change the question forms. He tried to guide and to explain a lot to lead the interviewees to the points."* ***LT3:*** *"A little if the interviewee wanted confirmation."* ***LT4:*** *"He didn"t help in the way of simplifying. He repeated and adjusted the questions but not to be easier. The third candidate didn"t get the question and the interviewer repeated it a few times but he didn"t make it easier. He just wanted to get the answer."* ***LT5:*** *"Yes, I do because there were some questions that the test-takers couldn"t answer or answered off the point and the interviewer tried to ask again."*

All linguistic/trained raters, except LT1 whose answer was *"no"*, seem to agree that the interlocutor would *"repeat the questions again"* when he wanted *"to lead the interviewees to the points"* (LT2) or when *"the interviewee wanted confirmation"* (LT3) or when *"the test-takers couldn"t answer or answered off the point"* (LT5). The interlocutor *"didn"t help in the way of simplifying". "He just wanted to get the answer"* (LT4).

Table 4.146 shows the linguistic/untrained raters"(LU) considerations concerning the candidates"age.

Table 4.146: The linguistic/untrained raters"(LU) considerations concerning the interviewers"accommodation

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
33. The interviewers/ interlocutors tried to help/accommodate the candidate during the test	x		x			x		x	x		***LU1:*** *"Yes, in the interview part because he didn"t push the candidates too much. He didn"tjust let it pass." "It was like when the interviewer realized that the candidate didn"t know more than that, he just gave up."* ***LU2:*** *"Yes, especially in the last part because the way the interviewer changed the questions to be easier for the candidates to be able to answer."* ***LU3:*** *"No, he didn"ttry to help, just tried to make them speak. It was like if they didn"t understand the questions, he repeated them or simplified them."* ***LU4:*** *"Yes maybe because he gave cues in answering the questions." "I don"t know if they*

	saw each other during the test. If they didn"t, okay, they didn"t help." "They might help by repeating the questions." ***LU5:*** *"A little." "Well, he didn"t help by leading to the answers. He kind of tried to restate the question to get the gist but not directly gave the answer." "It was rather facilitation."*

LU1 and LU2 seemed to have strong feelings that the interlocutor tried to help the candidates. LU1 stated that *"it was like when the interviewer realized that the candidate didn"t know more than that, he just gave up"* while LU2 detected that *"the interviewer changed the questions to be easier for the candidates to be able to answer."* LU4 and LU5 did not feel that strong about the interlocutor. For LU4 it was just that *"they might help by repeating the questions"* and *"it was rather facilitation"* for LU5 who thought that the interlocutor just *"kind of tried to restate the question to get the gist but not directly gave the answer"*. LU3 was the only linguistic/untrained rater who did not think so. The interlocutor *"didn"t try to help, just tried to make them speak"* in the opinion of LU3.

Table 4.147 shows the operational/trained raters" (OT) considerations concerning the candidates" age.

Table 4.147: The operational/trained raters" (OT) considerations concerning the interviewers" accommodation

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
33. The interviewers/interlocutors tried to		x	x			x	x		x		***OT1:*** *"No, I believe that the interviewer was very professional."*

help/accommodate the candidate during the test	***OT2:*** *"Yes." "The interviewer was trying to help by saying some keywords or just tried to lead him so he could get the answer."* ***OT3:*** *"Not really. No."* ***OT4:*** *"Yes. Because the interviewer knew that this had effect on the test-takers" career so it"s better to help each other." "The interviewer helped by rephrasing and clarifying the questions. He also spoke with slower speed."* ***OT5:*** *"Yes, he tried to simplify. Just a bit of help, I guess. It wouldn"t influence the answer. No."*

Among the operational/trained raters, OT1 and OT3 said firmly that the interlocutor did not help the candidates because *"the interviewer was very professional"* (OT1). However, OT2 had different point of view. *"The interviewer was trying to help by saying some keywords or just tried to lead him so he could get the answer"* was his answer. OT4 had similar idea that *"the interviewer helped by rephrasing and clarifying the questions"* and *"he also spoke with slower speed"* because *"the interviewer knew that this had effect on the test-takers" career so it"s better to help each other".* OT5 explained that the interlocutor *"tried to simplify"* which was *"just a bit of help"* but *"it wouldn"t influence the answer".*

Table 4.148 shows the operational/untrained raters" (OU) considerations concerning the candidates" age.

Table 4.148: The operational/untrained raters" (OU) considerations concerning the interviewers" accommodation

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
33. The interviewers/interlocutors tried to help/accommodate the candidate during the test		x	x			x	x		x		***OU1:*** *"No, he didn"t because it was the same pattern." "He did the same thing to every candidate. If they didn"tunderstand, he just repeated a little."* ***OU2:*** *"Yes, partly." "He tried to use easier terms or paraphrase."* ***OU3:*** *"I didn"t feel that."* ***OU4:*** *"Yes, he did because he tried to simplify his questions. He didn"t try to give the answers."* ***OU5:*** *"Yes, partly. Sometimes the candidates" answers were not direct to the point, the interviewer changed the question or repeated it or simplified it. He tried to lead the candidates to the get the answers which were straight to the point. Otherwise they would answer incorrectly or didn"t answer at all."*

OU1 and OU3 did not think that the interlocutor tried to help the candidates *"because it was the same pattern"* and *"he did the same thing to every candidate"* (OU1). On the other hand, OU 2 and OU5 said that it was *"partly"*. In the opinion of OU2, the interlocutor *"tried to use easier terms or paraphrase"* while *"he tried to lead the candidates to the get the answers which were*

straight to the point" for OU5. Even though OU4"s answer was *"yes",* he thought that the interlocutor *"tried to simplify his questions".* He emphasized that the interlocutor *"didn"t try to give the answers".*

Table 4.149 shows the linguistic/trained raters" (LT) considerations concerning the interviewers" speech simplification.

Table 4.149: The linguistic/trained raters" (LT) considerations concerning the interviewers" speech simplification

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
34. The interviewers/interlocutors attempted to simplify their speech to facilitate the candidates or to match the candidates" level of language		x	x		x			x	x		***LT1:*** *"No."* ***LT2:*** *"Yes, both simplified and explained."* ***LT3:*** *"Yes, when the interviewees" proficiency was inadequate."* ***LT4:*** *"No."* ***LT5:*** *"Yes, they simplified their questions."*

Two linguistic/trained raters (LT1 and LT4) did not think that the interviewers try to simplify their speech to facilitate the candidates or to match the candidates" level of language while the other three (LT2, LT3 and LT5) had an opposing idea.

Table 4.150 shows the linguistic/untrained raters" (LU) considerations concerning the interviewers" speech simplification.

Table 4.150: The linguistic/untrained raters‟ (LU) considerations concerning the interviewers‟ speech simplification

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
34. The interviewers/interlocutors attempted to simplify their speech to facilitate the candidates or to match the candidates‟ level of language	x		x		x		x		x		***LU1:*** *“Yes.” “Especially for the second candidate.”* ***LU2:*** *“Yes.” “He tried to exemplify his questions in the second part.”* ***LU3:*** *“Let me say he revised the questions for the second candidate.”* ***LU4:*** *“Yes, by simplifying or repeating the questions or slowing down the speech rate.”* ***LU5:*** *“Yes. During the interview, when the candidates didn‟t understand, the interviewer tried to explain the questions for them.”*

All linguistic/untrained raters concurred that they thought the interviewers try to simplify their speech to facilitate the candidates or to match the candidates‟ level of language.

Table 4.151 shows the operational/trained raters‟ (OT) considerations concerning the interviewers‟ speech simplification.

Table 4.151: The operational/trained raters‟ (OT) considerations concerning the interviewers‟ speech simplification

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
34. The interviewers/interlocutors attempted to simplify their speech to facilitate the candidates or to match the candidates‟ level of language		x		x	x		x		x		***OT1:*** *“No. however, what they did was … um…try to ask the questions more than once.” “Maybe give him a chance to answer the question.”* ***OT2:*** *“Umm…ah…no…maybe…his vocabulary stepped down. He could have used words … another Australian or American.” “Not all cases. I think I heard once or twice.”* ***OT3:*** *“Yes.”* ***OT4:*** *“Yes.”* ***OT5:*** *“Yes.”*

OT1 and OT2 said „no‟ whether they thought the interviewers tried to simplify their speech to facilitate the candidates or to match the candidates‟ level of language. However, they still explained as if they thought of it the other way around as *“however, what they did was … um…try to ask the questions more than once”* (OT1) and *“maybe…his vocabulary stepped down”* (OT2). The other three operational/trained raters (OT3, OT4 and OT5) just simply answered „yes‟ to this question.

Table 4.152 shows the operational/untrained raters‟ (OU) considerations concerning the interviewers‟ speech simplification.

Table 4.152: The operational/untrained raters‟ (OU) considerations concerning the interviewers‟ speech simplification

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
34. The interviewers/interlocutors attempted to simplify their speech to facilitate the candidates or to match the candidates‟ level of language	x		x			x	x		x		***OU1:*** *“He just rephrased a little.”* ***OU2:*** *“Yes. He tried to repeat or change the questions to help the candidate understand the questions.”* ***OU3:*** *“No, I don‟t think so. I think it was standard.” “He might explain a little. I don‟t think he simplified the language but he explained more. It was an additional explanation. Some but not much.”* ***OU4:*** *“Yes. It‟s normal in an interview. The interviewee has right to ask back to the interviewer if the question is unclear.”* ***OU5:*** *“Yes.”*

Three operational/untrained raters (OU2, OU4 and OU5) thought the interviewers tried to simplify their speech to facilitate the candidates or to match the candidates‟ level of language. OU1 did not clearly state his answer as „yes‟ or „no‟ but he said that the interviewer *“just rephrased a little”*. So it could be implied that his answer was „yes‟. Even though OU3‟s answer was „no‟ but he admitted that the interviewer *“might explain a little”*.

Table 4.153 shows the linguistic/trained raters‟ (LT) considerations concerning the interviewers‟ performance.

Table 4.153: The linguistic/trained raters‟ (LT) considerations concerning the interviewers‟ performance

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
35. The interviewers performed their jobs appropriately.	x		x		x		x		x		***LT1:*** *"Yes."* ***LT2:*** *"Quite okay."* ***LT3:*** *"Yes."* ***LT4:*** *"Yes."* ***LT5:*** *"Yes."*

Every rater in the linguistic/trained group agreed that the interviewers performed their jobs appropriately.

Table 4.154 shows the linguistic/untrained raters‟ (LU) considerations concerning the interviewers‟ performance.

Table 4.154: The linguistic/untrained raters‟ (LU) considerations concerning the interviewers‟ performance

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
35. The interviewers performed their jobs appropriately.	x		x		x		x		x		***LU1:*** *"Quite well."* ***LU2:*** *"I think it was okay."* ***LU3:*** *"Yes."*

	LU4: *"Yes."* ***LU5:*** *"Yes."*

All linguistic/untrained raters concurred that the interviewers performed their jobs appropriately.

Table 4.155 shows the operational/trained raters'' (OT) considerations concerning the interviewers'' performance.

Table 4.155: The operational/trained raters'' (OT) considerations concerning the interviewers'' performance

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	
35. The interviewers performed their jobs appropriately.	x		x		x		x		x		***OT1:*** *"Yes."* ***OT2:*** *"Yes, out of 100 I give him 95."* ***OT3:*** *"Yes."* ***OT4:*** *"Yes. Eight out of ten."* ***OT5:*** *"The interviewer? Um … yeah. They got a ratable speech sample. So I guess they did the job fine.*

The operational/trained were unanimous that the interviewers performed their jobs appropriately. OT5 also explained the reason why he justified that *"as long as you get a ratable speech sample then the interviewer has done his job"*

Table 4.156 shows the operational/untrained raters‟ (OU) considerations concerning the interviewers‟ performance.

Table 4.156: The operational/untrained raters‟ (OU) considerations concerning the interviewers‟ performance

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
35. The interviewers performed their jobs appropriately.	x		x		x		x		x		***OU1:*** *"Yes."* ***OU2:*** *"Yes."* ***OU3:*** *"Um ... it was okay but I think there should have been some variation. He might vary a little but within acceptable level."* ***OU4:*** *"Yes."* ***OU5:*** *"Yes."*

The operational/untrained raters‟ idea was uniform that the interviewers performed their jobs appropriately. OU3 seemed to have a little doubt that the interviewer should have *"vary the questions a little bit otherwise the latter test-takers would know the format"* but he still thought *"it was okay"*.

Table 4.157 shows the linguistic/trained raters‟ (LT)considerations concerning the candidates‟ age.

Table 4.157: The linguistic/trained raters‟ (LT) considerations concerning the candidates‟ age

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
36. Taking the candidates‟ age into considerations		x		x		x		x	x		***LT1:*** *“No. Age is no concern.”* ***LT2:*** *“No. (Laughter)* ***LT3:*** *“No.”* ***LT4:*** *“No.”* ***LT5:*** *“Yes, a little.” “It wasn‟t their „life age‟. It was their „job age‟ - their experience. Ones who have more „job age‟ should have more proficiency in some … what? … some criteria. But not all.” “I knew their „life age‟ because they mentioned it in the interview.”*

All linguistic/trained raters agreed that the candidates‟ age was not their concerns in their ratings. However, LT5 accepted that she was *“a little”* concerned about the candidates‟ „job age‟ or their experience because she thought that *“ones who have more „job age‟ should have more proficiency in some criteria”*.

Table 4.158 shows the linguistic/untrained raters‟ (LU) considerations concerning the candidates‟ age.

Table 4.158: The linguistic/untrained raters‟ (LU) considerations concerning the candidates‟ age

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
36. Taking the candidates‟ age into considerations	x			x		x		x		x	***LU1:*** *"No, I didn"t." "But when the candidates answered how long they had been flying, that means their experience in flying." "I partly considered that." "Yes, I did consider."* ***LU2:*** *"No." "A candidate said that he had been flying for 12 years. It made me think that he should have done better if he had that much experience." "But I didn"tthink about this while I was rating." "I still rated according to what I heard but I couldn"t help thinking that you had 12 years of experience and these were all you could do!"* ***LU3:*** *"No."* ***LU4:*** *"No."* **LU5:** "No."

All linguistic/untrained raters answered that they did not consider the candidates‟ age in their ratings. Though two raters (LT1 and LT2) admitted that she *"partly considered"* their experience (LT1) and it made LT2 thought that the candidate *"should have done better if he had that much experience"*.

Table 4.159 shows the operational/trained raters‟ (OT) considerations concerning the candidates‟ age.

Table 4.159: The operational/trained raters' (OT) considerations concerning the candidates' age

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
36. Taking the candidates' age into considerations		x		x		x		x		x	***OT1:*** *"No."* ***OT2:*** *"I didn't consider that because I've been here too long. I know how Thai pilots think. That's why."* ***OT3:*** *"No."* ***OT4:*** *"No."* ***OT5:*** *"No."*

All five operational/trained raters simply said that they did not consider the candidates' age in their ratings.

Table 4.160 shows the operational/untrained raters' (OU) considerations concerning the candidates' age.

Table 4.160: The operational/untrained raters' (OU) considerations concerning the candidates' age

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
36. Taking the candidates' age into considerations		x		x		x		x		x	***OU1:*** *"No."* ***OU2:*** *"No."* ***OU3:*** *"No."* ***OU4:*** *"No."*

	OU5: *"I didn't know their age."*

All five operational/untrained raters plainly said that they did not consider the candidates' age in their ratings. OU5 also clarified that was because he *"didn't know their age"*.

Table 4.161 shows the linguistic/trained raters' (LT) considerations concerning the candidates' gender.

Table 4.161: The linguistic/trained raters' (LT) considerations concerning the candidates' gender

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
37. Taking the candidates' gender into considerations		x		x		x		x		x	***LT1:*** *"No." (Laughter)* ***LT2:*** *"No." (Laughter)* ***LT3:*** *"No, but all candidates I've ever rated were all male."* ***LT4:*** *"No."* ***LT5:*** *"No."*

All five linguistic/trained raters simply said that they did not consider the candidates' gender in their ratings. LT3 added that all candidates she had ever rated *"were all male"*.

Table 4.162 shows the linguistic/untrained raters' (LU) considerations concerning the candidates' gender.

Table 4.162: The linguistic/untrained raters' (LU) considerations concerning the candidates' gender

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
37. Taking the candidates' gender into considerations		x		x		x		x		x	***LU1:*** *"No because I expected only male test takers."* ***LU2:*** *"No."* ***LU3:*** *"No."* ***LU4:*** *"No."* ***LU5:*** *"No."*

All five linguistic/untrained raters plainly said that they did not consider the candidates' gender in their ratings. LU1 added that she *"expected only male test takers."*

Table 4.163 shows the operational/trained raters' (OT) considerations concerning the candidates' gender.

Table 4.163: The operational/trained raters' (OT) considerations concerning the candidates' gender

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
37. Taking the candidates' gender into considerations		x		x		x		x		x	***OT1:*** *"No."* ***OT2:*** *"No. I always believe women can fly as well as men."*

											OT3: *"No."*
											OT4: *"No."*
											OT5: *"No."*

All five operational/trained raters said that they did not consider the candidates" gender in their ratings. OT2 also emphasized that he *"always"* believed that *"women can fly as well as men"*.

Table 4.164 shows the operational/untrained raters" (OU) considerations concerning the candidates" gender.

Table 4.164: The operational/untrained raters" (OU) considerations concerning the candidates" gender

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
37. Taking the candidates" gender into considerations		x		x		x		x		x	***OU1:*** *"No (Laughter)."*
											OU2: *"No."*
											OU3: *"No."*
											OU4: *"No."*
											OU5: *"No."*

All five operational/untrained raters merely said that they did not consider the candidates‟ gender in their ratings with no further explanation.

Table 4.165 shows the linguistic/trained raters‟ (LT) considerations concerning the candidates‟ global/overall attitudes.

Table 4.165: The linguistic/trained raters‟ (LT) considerations concerning the candidates‟ global/overall attitudes

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
38. Taking the candidates‟ global/ overall attitudes into considerations		x		x		x		x	x		***LT1:*** *“It‟s not an image that we create. It‟s the way they answer the questions. It‟s not the confidence in their voice but it‟s the confidence they possess in doing their jobs.”* ***LT2:*** *“Not at all.”* ***LT3:*** *“Not at all. We must not know them. We must not have any bias.”* ***LT4:*** *“No.”* ***LT5:*** *“Their experience should be in this category, shouldn‟t it? So, yes, I considered them too.”*

Almost all linguistic/trained raters (LT1, LT2, LT3 and LT4) said that they did not consider the candidates‟ overall attitudes except LT5 still confirmed that she considered their experience.

Table 4.166 shows the linguistic/untrained raters‟ (LU) considerations concerning the candidates‟ global/overall attitudes**.**

Table 4.166: The linguistic/untrained raters‟ (LU) considerations concerning the candidates‟ global/overall attitudes

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
38. Taking the candidates‟ global/ overall into considerations		x		x	x		x			x	***LU1:*** *"No. It's hard to consider that from listening only."* ***LU2:*** *"No."* ***LU3:*** *"No." "Well, may be the voice tone. Some had confident voice tone. It sounded relaxing but some sounded … nervous." "It might have some effect. When he sounded confident, I felt positive about him." "I felt that he shouldn"tget low score. Sometimes it made me forget about his mistakes."* ***LU4:*** *"No because I didn"t see them." "Well, maybe their confidence."* ***LU5:*** *"No." That"s because I didn"t see them in person. I just listened to their voice."*

All linguistic/untrained raters answered *"no"* in the beginning if they considered the candidates‟ overall attitudes. However, LU3 accepted that when the candidates sounded confident she *"felt positive about him"*. LU4 also admitted that it was *"maybe their confidence."*

Table 4.167 shows the operational/trained raters‟ (OT) considerations concerning the candidates‟ global/overall attitudes.

Table 4.167: The operational/trained raters‟ (OT) considerations concerning the candidates‟ global/overall attitudes

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
38. Taking the candidates‟ global/ overall into considerations		x		x		x		x		x	***OT1:*** *"No."* ***OT2:*** *"I know who all three of them are. No, I didn"t consider. I'm not biased." "When I was giving the scores, I wanted to sympathize but then (Laughter) otherwise they would have all passed."* ***OT3:*** *"No. I just looked at their English."* ***OT4:*** *"No." "I actually knew all of the test-takers. I had their faces in my mind but I didn"t take that into my consideration."* ***OT5:*** *"Attitudes?" "No, I wouldn"t.Just looking at the answers."*

The difference between the groups of the linguistic (both trained and untrained) raters and the operational/trained raters was that the latter group *"knew all of the test-takers"* but all of them said that they did not consider the overall attitude of the candidates in their ratings.

Table 4.168 shows the operational/untrained raters‟ (OU) considerations concerning the candidates‟ global/overall attitudes.

Table 4.168: The operational/untrained raters‟ (OU) considerations concerning the candidates‟ global/overall attitudes

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
38. Taking the candidates‟ global/ overall into considerations	x			x	x			x		x	***OU1:*** *"It has influence but it"s not alright." "It shouldn"t be done." "We shouldn"t see the candidates, just listen to their voice."* ***OU2:*** *"No because I didn"t really see them in persons."* ***OU3:*** *"I don"t know them." "Yes, a little. I gave the second candidate not a good score because I thought he covered up something. Cover up in the way that he didn"t really comprehend but he answered promptly. He tried to show that he was confident by answering right away. He replied quickly and prematurely to cover up his weak points. It might not be intention but his subconscious."* ***OU4:*** *"No because we didn"t rate them live. We just listened to their voices."* ***OU5:*** *"No because I didn"t see them in the flesh, just their voices."*

OU1 accepted that the overall attitude of the candidates *"has influence"* thought *"it"s not alright"*. OU3 also admitted that he considered it *"a little"* because he thought the candidate *"covered up something"*. The other three (OU2, OU4 and OU5) said that

they did not consider the overall attitude of the candidates in their ratings with the similar reason that they did not *"really see them in persons"* (OU2), did not *"rate them live"* (OU4) and did not *"see them in the flesh"*.

Table 4.169 shows if the linguistic/trained raters (LT) thought the candidates were nervous during testing**.**

Table 4.169: The linguistic/trained raters" (LT) thoughts if the candidates were nervous during testing

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
39. The candidates were nervous during testing.	x		x			x	x		x		***LT1:*** *"Yes, number three. Because he struggled to answer the questions and he was also confused with the sequences of events. He even admitted that he never experienced that kind of event before." "He was panicked."* ***LT2:*** *"Somewhat. Particularly the second and the third candidates. The first one was quite confident."* ***LT3:*** *"Umm…a little." "Let"s say no because it"s their job." "They might be nervous if they started with the interview part but they started with the radiotelephony part so they could concentrate on that. If they were, I didn't feel it."* ***LT4:*** *"No, not the first one. He was quite relaxed." "Perhaps a little for the second and the third guys."*

	LT5: *"Yes, some of them. Not the first guy but the other two."*

All linguistic/trained raters accepted that they thought at least one candidate was nervous to some extent. However, LT3 changed her answer from *"a little"* to *"I didn't feel it"* later.

Table 4.170 shows if the linguistic/untrained raters (LU) thought the candidates were nervous during testing**.**

Table 4.170: The linguistic/untrained raters" (LU) thoughts if the candidates were nervous during testing

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	
39. The candidates were nervous during testing.	x		x		x		x		x		***LU1:*** *"The first and the third candidates seemed to be alright but the second seemed to be the weakest."* ***LU2:*** *"Yes, all three, especially under tough situations."* ***LU3:*** *"Yes, in the first part."* ***LU4:*** *"Yes, the third because he said by himself that he was nervous." "I think the second and the third were nervous."* ***LU5:*** *"Yes. The third was the most."*

All linguistic/untrained raters accepted that they thought at least one candidate was nervous to some extent.

Table 4.171 shows if the operational/trained raters (OT) thought the candidates were nervous during testing.

Table 4.171: The operational/trained raters (OT) thoughts if the candidates were nervous during testing

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
39. The candidates were nervous during testing.		x		x		x		x	x		***OT1:*** *"Yes, the last two were nervous when they got into …um…abnormal situations, especially number three when he couldn"t express his idea about the future equipment, he began to get nervous and his speech began to go faster and made more mistakes in grammar and sentence structures."* ***OT2:*** *"I think all of them were nervous. The last one was the most nervous and the first one was the most at ease but they all had a short period of nervousness and then, oh! okay. This is what it"s all about."* ***OT3:*** *"Yes."* ***OT4:*** *"Yes."* ***OT5:*** *"The first one, I think, wasn"tsince he was a captain candidate, he had the experience, he had the knowledge of how to." "The second one he had*

	some when he faced with the non-normal situations, he"s quite nervous. Also the third but not as much as the second."

All operational/trained raters accepted that they thought at least one candidate was nervous to some extent. OT1, OT2 and OT5 seemed to be of the same opinion that the first candidate was not or the least nervous while the third was the most nervous, though OT5 had the different idea that the second candidate was more nervous than the third.

Table 4.172 shows if the operational/untrained raters (OU) thought the candidates were nervous during testing.

Table 4.172: The operational/untrained raters (OU) thoughts if the candidates were nervous during testing

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	
39. The candidates were nervous during testing.		x	x		x		x		x		***OU1:*** *"No, I don"t think so."* ***OU2:*** *"Yes."* ***OU3:*** *"Yes, I think so." "I might be biased against the first candidate a little because I realized later that he was a captain candidate. He was the least nervous candidate."* ***OU4:*** *"Yes, all of them."* ***OU5:*** *"Yes, the second guy."*

OU1 was the only operational/untrained rater who said that he did not think any candidate was nervous. The other four raters (OU2, OU3, OU4 and OU5) thought that at least one candidate was nervous to some extent. OU3 also accepted that he *"might be biased against the first candidate a little"* because he *"realized later that he was a captain candidate"*.

Table 4.173 shows if the linguistic/trained raters (LT) sympathized for the candidates" nervousness in their ratings.

Table 4.173: The linguistic/trained raters" (LT) sympathy for the candidates" nervousness in their ratings

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
40. Sympathized for the candidates" nervousness in their ratings		x		x	N.a.	N.a.		x	x		***LT1:*** *"No. I had no pity on him because he couldn"t answer by himself."* ***LT2:*** *"No." (Laughter)* ***LT3:*** *(Not applicable because she said she did not feel if any candidate was nervous.)* ***LT4:*** *"No because they might be nervous but they knew what they were doing."* ***LT5:*** *"Yes, a little. Not in all aspects because in the situation like this their proficiency might drop a little. Their actual proficiency might not be that low. But if they did something repeatedly, it was their true problem. It wasn"t because of their nervousness. I sympathized with them a little for that."*

Only one linguistic/trained rater (LT5) said that she sympathized for the candidates‟ nervousness in her rating because *"in the situation like this their proficiency might drop a little"*. The other three raters said that they did not sympathize for the candidates‟ nervousness in their ratings. LT3 was the only one who answered in the previous question that she did not feel if any candidate was nervous (N.a. = Not applicable).

Table 4.174 shows if the linguistic/untrained raters (LU) sympathized for the candidates‟ nervousness in their ratings.

Table 4.174: The linguistic/untrained raters‟ (LU) sympathy for the candidates‟ nervousness in their ratings

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
40. Sympathized for the candidates‟ nervousness in their ratings	x			x		x		x		x	***LU1:*** *"Honestly, yes."* ***LU2:*** *"No."* ***LU3:*** *"No, not sympathized."* ***LU4:*** *"No."* ***LU5:*** *"No. I don"t think so because when people get nervous, they"ll lose their fluency. We can"t sympathize with them for this."*

LU1 was the only linguistic/untrained rater who accepted that she sympathized for the candidates‟ nervousness in her rating. The other four (LU2, LU3, LU4 and LU5) said that they did not sympathize for that in their ratings.

Table 4.175 shows if the operational/trained raters (OT) sympathized for the candidates‟ nervousness in their ratings**.**

Table 4.175: The operational/trained raters‟ (OT) sympathy for the candidates‟ nervousness in their ratings

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
40. Sympathized for the candidates‟ nervousness in their ratings		x	x			x	x			x	***OT1:*** *“No. I just rated him.”* ***OT2:*** *“Maybe for the first few questions.”* ***OT3:*** *“No.”* ***OT4:*** *“Four out of ten.”* ***OT5:*** *“No. He gets what he deserves.”*

Three operational/trained (OT1, OT3 and OT5) said that they did not sympathize for the candidates‟ nervousness in their ratings. OT3 stated that he „partly‟ sympathized for that to a certain extent as *“for the first few questions”* while OT4 also „slightly‟ sympathized as *“four out of ten”*.

Table 4.176 shows if the operational/untrained raters (OU) sympathized for the candidates‟ nervousness in their ratings.

Table 4.176: The operational/untrained raters‟ (OU) sympathy for the candidates‟ nervousness in their ratings

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
40. Sympathized for the	N.a.	N.a.		x		x		x		x	***OU1:*** *(Not applicable because he said he did not*

candidates‟ nervousness in their ratings	*feel if any candidate was nervous.)* ***OU2:*** *"No."* ***OU3:*** *"No, I didn"t."* ***OU4:*** *"No, I don"t think so."* ***OU5:*** *"I didn"t sympathize because he was nervous but … what should I say? … I tried to see if he understood what the interviewer said." "I just realized that he was nervous."*

All four operational/untrained raters(OU2, OU3, OU4 and OU5) , except OU1 who was the only one who answered in the previous question that he did not feel if any candidate was nervous, said that they did not sympathize for the candidates‟ nervousness in their ratings. OU5 also clarified that he *"just realized"* that they were nervous.

Table 4.177 shows if the linguistic/trained raters (LT) compared a candidate with the others.

Table 4.177: The linguistic/trained raters‟ (LT) comparison of a candidate with the others

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
41. Comparing a candidate with the others		x		x		x		x	x		***LT1:*** *"No. Not during rating." "Because I paid more attention on the criteria and the weaknesses made by the candidates which usually occurred."* ***LT2:*** *"No." "If we compare the best candidate today with the best yesterday, it would sway our standard. So I don"t compare, just with the scales."*

LT3: *"No." "I didn"t compare each individual but ...for example this guy got „four," in my mind I put „four"as a benchmark. So when I gave someone a „four," he must be okay." "I compared in terms of the scores because I put „four"as a benchmark." "I have „four"as a benchmark in my mind. If someone gets „four"then the picture of „four" will be clearer. He gets „four"according to the laid down criteria. Candidates who are assessed after this first gut will be compared. But other levels will not be compared. Not at all. Therefore we have to put level four as a benchmark first." "The picture of level four in the table will be clearer and that can be used as a benchmark but not other levels. This is my personal technique."*

LT4: *"No. They couldn"t be compared because their proficiency showed."*

LT5: *"Yes, I did." "I put the first guy as a benchmark. If the latter guy did better, he"d get better score. If he did worse, he"d get lower score." "There are two standards for benchmarking. The first one is the one which is laid down by ICAO scales. I compared those three candidates with this standard. However, when I grouped these three guys, I looked how they performed. It"s like having a ruler then we put another one and other two to compare between the first one and also between them." "I compared the first guy with the ICAO scales. After that I compared the second guy with both ICAO scales and the first one."*

Four out of five linguistic/trained raters (LT1, LT2, LT3 and LT4) said that they did not compare the candidate with other candidates in their ratings. LT3 said that she did not compare each individual but if she put a candidate as „four" *"then the picture of „four" will be clearer"*. LT5 was the only one in this group who admitted that she did. She *"put the first guy as a benchmark"* after comparing *"the first guy with the ICAO scales"* then she *"compared the second guy with both ICAO scales and the first one"*.

Table 4.178 shows if the linguistic/untrained raters (LU) compared a candidate with the others.

Table 4.178: The linguistic/untrained raters" (LU) comparison of a candidate with the others

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
41. Comparing a candidate with the others		x	x		x		x		x		***LU1:*** *"Because there were just three of them and I never did this before, so I didn"t."* ***LU2:*** *"Yes because I just finished with the first candidate, then the second one." "And because when I rate many people, I usually put the first one as a benchmark. Then I compare the latter ones with the first if they are better or worse. Sometimes I have to go back and change the score I gave the first candidate if I feel that the latter does better or worse."* ***LU3:*** *"Um ... I had that comparative feeling but during rating ... well, it was quite clear that it was like „four," „thræ","two".There were two candidates. I was giving one guy „thræ plus" and another „four". I compared the first guy with the third but I didn"tcompare the second because*

he obviously lagged behind. Their performance in the first part was similar but it was obvious that the first guy was more fluent in the second part. He was more confident in the interview than the third guy." "I compared so the third guy got the „plus". Their performance was not equivalent so he got the „plus"."

LU4: *"Yes, I did." (After a while) "I change my answer. No, I didn"t compare." "Actually I wanted to give the second candidate lower than „four"in some points but they didn"t interfere with his jobs. However, he had some problems in the interview that he didn"t understand some questions and his answers were not straight to the point. But I gave him for the overall." "He could do his job because he could use the correct phraseologies but he had problems with the interview. You know what I mean? Actually it should be divided between each part. He could handle the part concerning his job though he might pronounce with difficulties such as the word „turbulence" but he could operate his job. The ATC could understand him. The interview is another issue. If I gave him „thræ"in comprehension, he"d fail. I knew from the manual that „thræ" means „fail"."" "Raters have right to compare by intuition. Raters intuitively compare. But in theory, they shouldn"t compare." "In reality, I compared." "I compared in terms of their fluency, their pronunciation." "I put the first candidate as a benchmark. If the second guy did better, he"d get higher score. If he did worse, he"d get lower score." "It also depends on the rubric too but I certainly compared."*

LU5: *"Yes, I did." "As I said, I didn"t have any*

	background when rating the first candidate. If raters have some background, at least they may not compare because they have their own background to judge if the candidate is good or not. But for me, I admit that I didn"t get anything at all. I just looked at the overall. When I first listened to the first candidate, I thought he wasn"t good but after listening to the third I had to change the first"s score because I realized that „Hey! He was good"." "After listening to the second candidate and compared him with the first one, I realized that the first guy was good." "In the beginning I had no idea if a pilot speaking like that was okay or not. I just felt that it should have been better. But after listening to the second guy, I told myself „Hey! He was better." Raters who have background should know instantly that the first guy was good. But I didn"t have such background, so I had to compare."

The majority of the linguistic/untrained raters (LU2, LU3, LU4 and LU5) accepted that they compared the candidate with the others. LU2 said that when she rated many people, she *"usually put the first one as a benchmark"*. Then, she *"compares the latter ones with the first if they are better or worse"*. LU4 changed his answer twice from *"yes, I did"* to *"no, I didn"t compare"* and then to *"I certainly compared"* in the end. LU4 also added that *"raters have right to compare by intuition"* and *"raters intuitively compare"*. He finally concluded that *"but in theory, they shouldn"t compare"*; on the contrary, he admitted that *"in reality, I compared."* LU5 said that he compared because he *"didn"t have any background when rating the first candidate"*. He *"realized that the first guy was good" "after listening to the second candidate and compared him with the first one"*. LU1 was the only one who said that *"because there were just three of them"* and she *"never did this before"* so she did not compare.

Table 4.179 shows if the operational/trained raters (OT) compared a candidate with the others.

Table 4.179: The operational/trained raters" (OT) comparison of a candidate with the others

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
41. Comparing a candidate with the others		x		x	x		x		x		***OT1:*** *"No. I didn"t. I just rate in each section and I just grade in that section. Look at overall grade. Give them the one that would be the average." "In the training that we had, they said that not to compare. Because of the training I had, I could be able to separate each of the samples."* ***OT2:*** *"No because it"s not like that. So each one I did kind of individual."* ***OT3:*** *"Yes." "You can"t help that." "(Laughter) Because you kind of ... it"s like a cross check. You gave him the grade for number one. Then you get number two, you grade anything. It definitely wasn"t number one. So we compare it. Then you think, yeah okay, that"s right. And then you get number three, he"s better than number two, he"s better than number one. Okay, so I gave ... you know perhaps a better ... not a total score but score in some other parts and you just like cross check to verify what ... what grade I was giving was correct." "For example, you gave number one „four" and you listened to number two, you felt that this guy is worse than the first one. So number two would get not better than „thræ", not better than „four" because number one gets only „four". So number two must be something lower than or not more than „four"".* ***OT4:*** *"Yes because it"s natural. It happened automatically. It might be because I listened to them consecutively." "I*

	may not compare them if I rate one test-taker and then have a brake for a period of time before rating the next guys. If the first two guys are too good and the third one is just average, he may look bad or worse than he actually is in this set of test-takers." ***OT5:*** *"I wouldn"t say number one is better than number two. I would say he has more experience in his answers but actually being more fluent I"d say the third is more fluent in answering the questions. His English is better than the first and the second." "I would compare." "First of all you have standard rating of your own. Your own standard rating, right? What this guy needs to get a ,,four"? What this guy needs to get a ,,five"? What this guy needs to get a ,,six"? Since you"ve done the first one, you know that he only gets a standard here, right? Then you listen to the second one ... I mean you wouldn"t really compare. You would judge according to the rating scales you have but you would just maybe see if ..." "I wouldn"tcompare these two. I would compare to the scales I have." "You would compare with the rating scales."*

Two operational/trained raters (OT3 and OT4) admitted that they compared the candidate with the others. OT3 said it *"can"t help that"* while it was *"because it"snatural"* for OT4. The other three (OT1 and OT2) said that they did not do so. OT5"s answer was a little confused. He said *"I would compare"* once when he described the third candidate as *"his English is better than the first and the second"* then he said in the end that *"I wouldn"tcompare these two"* and *"I would compare to the scales I have".* This might be because he realized from his rater training that he should not compare a candidate with others.

Table 4.180 shows if the operational/untrained raters (OU) compared a candidate with the others.

Table 4.180: The operational/untrained raters" (OU) comparison of a candidate with the others

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
41. Comparing a candidate with the others	x			x	x		x		x		***OU1:*** *"Yes because we need a benchmark in our mind so we can compare." "After listening to all three, I knew who was the best, who was the second, and who was the worst." "I think I used the best as a benchmark." "Today it was the first candidate." "I looked at the overall performance so their scores might be the same but I compared in each criterion. For example, if the first guy"spronunciation was better and he got ,,five', the second guy who was worse wouldn"t get ,,five"."* ***OU2:*** *"No."* ***OU3:*** *"Yes because I can"t help it. It's human nature. Though they"re rated on another day, I still compare. Not much but still think of it." "But I also read the ICAO scales what level four is."* ***OU4:*** *"Yes." "No, I didn"t compare among them. I compared each of them with me." "I compared in the way that if I were the interviewer, how I would answer that question. I might answer like them or I might not understand the question like them or I would be excited like them when facing those situations." "If they did better than me, … (silent)." "I put myself as a benchmark. Today I thought of myself as a ,,level six" because I'm confident that I"m better than all of them. Then I compared them with me." "As a matter of fact, if I'm ,,six' I must be ,,six"*

	everyday, not just today." (Laughed) "I didn"t really base everything on me. I used my experience that I have to consider what level I should be. Do you understand what I mean? Among all pilots in Thai airways, I'm not the best but I can …" "What I mean is that if they do better than me, I can give them ,,six" but that"s it because it"s the highest level. No matter how better they are than me." ***OU5:*** *"Yes, somewhat. I compared them but not in terms of their scores. For example, I gave the first guy a certain score and the second guy did worse, I tried to see if he was worse and how much. Was he that worse that I had to award lower score or just a little worse but still acceptable to be in the same level? I compares in this sense. I put the first guy as a benchmark because I listened to him first." "It depends on the ability of the latter guys if they were better or worse." "I might look back and see if I gave the first guy too high or too low." "I still based the first guy on the ICAO guideline." "I tried to award the scores based on the guideline that I think it should be." "It's impossible to say I didn"t compare. It's human nature or you have to do the brainwash after each listening."*

Four operational/untrained raters (OU1, OU3, OU4 and OU5) accepted that they compared the candidate with the others. OU1 gave the reason as *"because we need a benchmark in our mind so we can compare"* while it was *"human nature"* for OU3. OU4 had an odd idea that he did not compare among the candidates but he compared them with himself. OU5 *"put the first guy as a benchmark"* because he *"listened to him first"*. OU2 was the only rater in this category who said he did not compare.

Table 4.181 shows the linguistic/trained raters" (LT) degrees of familiarity with the ICAO language proficiency rating scale.

Table 4.181: The linguistic/trained raters" (LT) degrees of familiarity with the ICAO language proficiency rating scale

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
42. Degrees of familiarity	x		x		x		x		x		***LT1:*** *"Very much."* ***LT2:*** *"Very much."* ***LT3:*** *"Very much."* ***LT4:*** *"Very much."* ***LT5:*** *"Pretty much."*

All linguistic/trained raters considered their familiarity with the ICAO language proficiency rating scale as *"very much"* (LT1, LT2, LT3 and LT4) and *"pretty much"* (LT5).

Table 4.182 shows the linguistic/untrained raters" (LU) degrees of familiarity with the ICAO language proficiency rating scale.

Table 4.182: The linguistic/untrained raters" (LU) degrees of familiarity with the ICAO language proficiency rating scale

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
42. Degrees of familiarity		x	x			x		x		x	***LU1:*** *"Not at all. This is my first time."* ***LU2:*** *"Not much."* ***LU3:*** *"Not at all. I've seen other kind of rating scale*

	before but this is the first time I see this scale." ***LU4:*** *"Not at all. This is my first time."* ***LU5:*** *"None. This is my first time."*

Almost all linguistic/untrained raters (LU1, LU3, LU4 and LU5) admitted that they were *"not"* familiar with the ICAO language proficiency rating scale *"at all"*. LU3 said that she had seen other kinds of rating scales before but it was her first time for this scale. LU2 was the only one who described her degree of familiarity as *"not much"*.

Table 4.183 shows the operational/trained raters" (OT) degrees of familiarity with the ICAO language proficiency rating scale.

Table 4.183: The operational/trained raters" (OT) degrees of familiarity with the ICAO language proficiency rating scale

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
42. Degrees of familiarity	x		x		x		x		x		***OT1:*** *"I would say not an expert in anyway."* ***OT2:*** *"Very much."* ***OT3:*** *"Let"s say„very much'". .* ***OT4:*** *"If I'm compared with other pilots in Thai Airways, I'm very familiar but if compared with those instructors in the language department, I'm in average." "It's four out of five."* ***OT5:*** *"I would say, from the scale one to ten, eight."*

Two operational/trained raters (OT2 and OT3) considered their familiarity with the ICAO language proficiency rating scale as *"very much"*. OT1 was rather modest to say that he was *"not an expert in anyway"* while OT4 and OT5 considered their familiarity as *"four out of five"* and *"eight"* out of ten respectively.

Table 4.184 shows the operational/untrained raters‟ (OU) degrees of familiarity with the ICAO language proficiency rating scale.

Table 4.184: The operational/untrained raters‟ (OU) degrees of familiarity with the ICAO language proficiency rating scale

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
42. Degrees of familiarity	x		x		x		x		x		***OU1:*** *"Not much."* ***OU2:*** *"A little."* ***OU3:*** *"Not much."* ***OU4:*** *"Not so familiar." "Today is the first time I see these details. I knew about the levels but never saw these descriptors. Yet this kind of rating is like the IELTS but IELTS has nine levels. I read IELTS descriptors but these ICAO descriptors are longer. It's like dividing IELTS into six levels instead of nine." "I'm familiar with this kind of rating because I used to study."* ***OU5:*** *"Moderately."*

The raters in the operational/untrained group considered their familiarity with the ICAO language proficiency rating scale as *"not much"* (OU1 and OU3), *"a little"* (OU2), *"not so familiar"* (OU4) and *"moderately"* (OU5). OU4 added that he *"knew about the levels but never saw these descriptors".* However, he said that he studied the IELTS scale before so he was familiar with this kind of rating but not particularly the ICAO scale.

Table 4.185 shows the linguistic/trained raters" (LT) interpretation of the ICAO scale descriptors.

Table 4.185: The linguistic/trained raters" (LT) interpretation of the ICAO scale descriptors

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Ql.	Qt.	Ql.	Qt.	Ql.	Qt.	Ql.	Qt.	Ql.	Qt.	
43. & 44. Descriptor interpretation i.e. qualitatively (Ql.) or quantitatively (Qt.)	x	x	x		x	x	x			x	***LT1:*** *(Long thinking and very hesitant)* ***LT1:*** *"Both." "No, I didn"t count." "It"s …um… the …" "It"s the estimation from my experience." "I put ,,four" as a standard and if he"s better than ,,four," he"ll get ,,five"." "Because I know that ,,four" is a cut-off core, I put ,,four" as a standard." "For example, ,,four" for pronunciation means mostly understandable, acceptable. I can"t explain more." "I don"t say it"s a feeling, it"s more like an experience from seeing a lot, seeing for many times."* ***LT2:*** *"The problem arises when we face with someone who is in between. ,,Almost never" is actually close to ,,non-existence". The proportion of the frequency is very little. Both ,,almost never" and ,,rarely" do not interfere with ,,ease of understanding" or very little." "I didn"t count." "I used my experience." "After we"ve rated for a while, we know*

that this is „level six", this is „level five"." "There is quite a difference between them. It can't be said concretely. It"s more like an abstract."

LT2: *"I do it qualitatively by using my knowledge and experience which I share with my colleagues." "No, I didn"t count because it"s not mathematics." "Language can be varied."*

LT3: *"It shows from the numbers of mistakes, numbers and quality. I didn"t count. It's how frequent we jot down. If he repeatedly makes the same mistakes, it is often." "It also depends on the quality. If that mistake is a minor one, even though it happens often but it does not interfere with the meaning, I'll overlook it."*

LT3: *"Both, as I said."*

LT4: *(Long pause) "It depends on the speech sample. If it"s short, we wouldn"t see the difference. The speech sample must be appropriately long."*

LT4: *"Qualitatively. „Usually" is more than „frequently". I measure them in terms of the meaning, to see if the meaning is incorrect, to judge after listening if the meaning is alright. For example, if a Thai says this to a foreigner, does it cause communication breakdown or just misunderstanding?" "„Sometimes" is in the middle while „usually" happens regularly." "If he tells a story without using past tense in the whole story, this is „usually"." "I counted but I didn"t use it in terms of the numbers." "It's a combination of qualitative and quantitative."*

LT5: *"I put them in percentage because it can"t be …" "It"s not an exact number." "For „sometimes", it"s in the middle so it"s around 50% with plus and minus 5%." "For „rarely", I added 10 to 15% with also plus and minus 5%." "It"s 60% for „frequently" and 70% for „usually"."*

LT5: *"Quantitatively." "I counted the frequency as much as I could. But it"s not exactly, just roughly."*

Each linguistic/trained rater seemed to show many pauses and, even, long pauses including some hesitations when being asked to explain how she interpreted the ICAO descriptors. LT1 paused for quite a long while and was very hesitant to speak up that the researcher had to continue with the next question. LT2 and LT3 showed differences in terms of the frequency though both of them said that they did not count the number of the mistakes made by the candidates but LT2 said that she used her experience. She explained that *"it can"t be said concretely"* and *"it"s more like an abstract"* while LT3 said that she considered the numbers and quality of the mistakes. LT4 did not give a clear explanation for her interpretation. After a long pause, she just said *"it depends on the speech sample"* and *"the speech sample must be appropriately long"*. LT5 was the only rater in this category who put her interpretation in terms of percentage, though it was not an exact number.

Two linguistic/trained raters (LT2 and LT4) said that they considered the ICAO descriptors qualitatively while the other two (LT1 and LT3) said they did them by using both quantitative and qualitative measures. However, both of them overtly stated that they *"didn"t count"*. LT1 also added that it was not a feeling. *"It"s more like an experience from seeing a lot, seeing for many times",* she said. The only rater who clearly claimed to do it quantitatively by counting the frequency was LT5.

Table 4.186 shows the linguistic/untrained raters" (LU) interpretation of the ICAO scale descriptors.

Table 4.186: The linguistic/untrained raters" (LU) interpretation of the ICAO scale descriptors

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Ql.	Qt.	Ql.	Qt.	Ql.	Qt.	Ql.	Qt.	Ql.	Qt.	
43. & 44. Descriptor interpretation i.e. qualitatively (Ql.) or quantitatively (Qt.)	x			x	x		x		x		***LU1:*** *(After thinking carefully for a while) "It's quite difficult to express."* ***LU1:*** *"Qualitatively." "It may be a combination of both." "However, in my rating I used my feeling."* ***LU2:*** *"I do the same as when I assess writing. I put a number in my mind how many mistakes to be as „rarely"."* ***LU2:*** *"I do it quantitatively because I'm familiar to it." "I put marks on a paper, for example one mistake one mark. After finishing, I count those marks and consider the final score." "In speak assessment like this, I roughly average them from those marks."* ***LU3:*** *"For „frequently" it interferes almost all the time. His pronunciation was hardly understandable. For „sometimes" his pronunciation was generally understandable, just occasionally unintelligible." "I didn"t count the frequency, just used my feeling." "But for „six" it"s thoroughly intelligible." "It's hard to explain." "It's a feeling."* ***LU3:*** *"Qualitatively." "I looked at the overall pictures."*

"I used my feeling." "For example, in terms of fluency I looked at how often they stopped speaking. When they answered each question, how often they paused."

LU4: *"This is an interesting question." "It's hard to define." "„Rarely" for comprehension or fluency is more fluent than other candidates." "I didn"t count." "I considered from the overall picture, from the beginning to the end then I analyzed what level they should be in." "I analyzed by listening and looking at the notes I took where the problems were." "Each pilot had some similar and some different problems. I looked at my notes."*

LU4: *"Qualitatively by using my judgment." "I used my experience and feeling." "Raters must be trained to be able to judge this correctly."*

LU5: *"Umm… these terms are quite difficult for me to understand. Frankly, I read them but I didn"t understand their meanings. For example, it says that for „always" they should be able to speak in unexpected circumstances. I had no idea what „unexpected circumstances" mean in aviation. Therefore, I couldn"t rate if it was „always" or not. As a teacher, I just see if they could communicate by using the language in a way which „always" initiates … it means if they could exchange words with the interviewer with confidence. I just looked at this. I couldn"t judge if they could „always" use the language in those unexpected situations. I couldn"t say if they „always" used the language to solve the problems." "I used what so called „my general comprehension".For example, level four … to*

	me, I dared not give „five" or „six" because I don"t know how good pilots are to get level five or six. I have no idea at all." "I don"t know their discreteness." "Psychologically, when we rate something we are unfamiliar with, we tend to award them in-between, not too high, not too low. Because we rate them too low, well, perhaps they are actually good and that"s all we give them? On the other hand, if we give them „five" or „six", we would doubt if they"re really good. So I"d better give „four"." *"In summary, I can"t explain those terms. I just looked at the overall performance. „Four", well, it"s acceptable. That"s all I considered." "It takes more time to thoroughly study these rubrics, not just five minutes. You need to attend a workshop to study these in details. It must be clear. The rubrics are clear and raters must clearly understand them too. I just looked at the overall performance. I didn"t consider them in details." "It might also be a cultural Sub-themes. We, Thai people don"t like humiliating the others. We don"t want to hurt the others" feelings. That"s the case of low rating. But if we rate them too high, it"s sort of doing too much. It"s kind of trying to be neutral."* ***LU5:*** *"Qualitatively."*

Three linguistic/untrained raters accepted that it was either *"quite difficult to express"* (LU1) or *"hard to explain"* (LU3) or *"hard to define"* (LU4) to interpret the ICAO descriptors. LU5 admitted that he *"didn"t understand their meanings"*. That was why he could not *"explain those terms"*. He said that he *"had no idea what „unexpected circumstances" mean in aviation"* so he *"couldn"t rate if it was „always" or not"*. LU5 also confessed that he *"dared not give „five" or „six"* because he did not know *"how good pilots*

are to get level five or six". In addition, he mentioned that *"psychologically, when we rate something we are unfamiliar with, we tend to award them in-between, not too high, not too low*". This implies that LU5"s ratings were influenced by the effect of central tendency error. LU2 was the only rater in this batch who said that she *"put a number"* in her mind to decide how many mistakes should have been rated as „rarely"."

Four linguistic/untrained raters (LU1, LU3, LU4 and LU5) stated that they considered the ICAO descriptors qualitatively. LU1 and LU3 said that they used their feelings while LU4 used both her feeling and her experience. LU5 did not explained in detail, just briefly said that he did it qualitatively. The only rater in this group who declared that she did it quantitatively was LU5.

Table 4.187 shows the operational/trained raters" (OT) interpretation of the ICAO scale descriptors.

Table 4.187: The operational/trained raters" (OT) interpretation of the ICAO scale descriptors

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Ql.	Qt.	Ql.	Qt.	Ql.	Qt.	Ql.	Qt.	Ql.	Qt.	
43. & 44. Descriptor interpretation i.e. qualitatively (Ql.) or quantitatively (Qt.)	x	x	x	x	x			x	x		***OT1:*** *"I used my experience in my training and in real life."* ***OT1:*** *"I tried to use even judgment. Well … it"s difficult." "I didn"t count the errors." "However, if the subject made multiple, multiple errors in one section of the test, then it becomes the thing … say… how many times now. But if you look at the whole interview, then it"s different." "That"s why I say it"s so difficult but I try to do it equally."* ***OT2:*** *"For „almost never", in my mind I would give it one*

in fifty sentences that you speak. I didn"t actually count but I kind of make a percentage in my head. But the only problem with that is if I miss something, instead of having two mistakes in fifty, it could have been four mistakes in fifty. If I miss certain parts whether I speak about some other criteria or whatever, that could be the difference between a „rarely" and „only sometimes". So if ... how I do, I put it in percentage." "I didn"t actually count. I jotted down the notes of the mistakes." "For „almost never", he would probably get it like 95% of ...he would not make mistakes." ""Rarely" ...let"s say seven to ten. „Only sometimes" ... not more than fifteen percent. „Frequently" would be ...ah...let"s say fifty percent. „Usually" ...ah...above fifty percent (laughter)."

OT2: *"After I listened to whole thing, I would compare to my notes then I can see how many mistakes that he did on what. On my notes I"d write each criterion and I"d have the words that he did wrong or the phrases that he put wrong, the structure, his vocabulary, like „bomb blasting" he couldn"t find the word „explosive" or „explosion". Okay. I"d write those things out. I didn"t take each ... I didn"t do that. And then I just based that on my notes, on what I heard."*

OT3: *(Sigh). "It"s very similar. The thing is that you can"t ... it"s not black and white. It"s not what you want it to. It"s how each person would interpret it. I mean you can get the system work, you can give totally different scores because they interpret „only rarely", „every now and then", „not very often", whatever. But this is the way I interpreted it. So this is my score. This is me." "Okay, maybe ..."*

OT3: *"Qualitatively. I definitely didn"t count it." "I took it as a whole. The first one I went through all these things. I think hold on, hold on, overall he isn"t that bad. Okay, he made mistakes, you know, just a bit but overall he"s okay. Overall is okay. So that"s why I stopped counting. I want more for quality and ... what, you know, everybody got his own mean for level four grade, you know, inside. And this is what he interprets level four and then when he comes to it, he listens to the whole speech sample and then first of all he thinks „do they match with his own level four interpretation? If one, alright, may be number two, I listened to it, well that"s definitely not level four, so I went back to the paper saying what is level four, what is level three, what is level five and then even though it says „only rarely", „every now and then", whatever, I didn"t count. I did it as overall quality check." "If you count, I find it tends to give a lower score." "Yeah, it"s hard to explain. It"s something like know what my level four threshold is. I know it"s like as a sim instructor, you"d say, okay, this is what you consider „pass", this is what you consider „fail". Then you look at the paper think, okay, now where could you put this one in but you got your own kind of ... integrity score. This is my level four. Did he meet it? Yes. Okay. Did he meet on every single one? I"m not sure but he met a certain level."*

OT4: *"It"s very in details, very close to each other. This was the topic that we discussed in our rater training course." "„Almost never" is very very little while „rarely" is the next level which is more than „almost never"." "I keep listening and see if he has a tendency to make the*

same thing repeatedly, this is „frequently".” “I took notes when it happened the first time, the second time, the third time. For „almost never" I took it as happened only once in the entire interview. Twice for „rarely". Three to four times for „only sometimes".” “It also depends on the importance and how often they say it. For example the „r" and „l" sounds, they might mispronounce because of the slip of the tongue. But those who mispronounce it by nature will say it regularly.” “For „frequently" it happens very often, may be five or six times. I consider „usually" the same as „frequently". They are in grey area which is hard to differentiate.” “Level three or two doesn"t matter. The cut-off score is level four.”

OT4: *“Quantitatively.”*

OT5: *“Well, I mean if he never makes any of the mistakes then I guess he"s perfect. No mistakes at all. No or very few. I mean a native speaker sometimes makes mistakes, right? Sometimes. But he makes very little. I wouldn"t be able to say how to judge those adverbs but you could see from the frequency ... from almost ... from very few that ...” “I don"t actually look at these adverbs, right? You look at so many criteria. You wouldn"t look at pronunciation only. You"d look at his speech sample to see how fluent he is. You would look at other criteria also.” “It"s a difficult question.” “I would see how frequent you make mistakes, I guess.” “I would have to count. I did that.” (Laughter and looked uncomfortable) “I couldn"t tell you the exact times. I would just see how many times, I guess. He makes ten out of ... how long his speech ... half*

an hour? He spoke like hundred twenty words per minute and he made five mistakes, it could be „almost never", may be, I don"t know. (Sighed) (Looked and sounded very uncomfortable) "I wouldn"t count exactly. You would look at the overall, looked at other criteria also. Not just concentrate on ... just one criterion."

OT5: *"I would rather base on qualitatively." "I mean half an hour for one person who can speak like a thousand words and half an hour for another person who can speak only five hundred words, right? So if you would count this guy one for one thousand words and this guy one for five hundred words then you couldn"t make it quantitatively, right? One for a thousand is that for „few" or „almost never"? Or one for five hundred is that „almost never"? Is that for „few"? I don"t know. So you base it on quality, I guess, not quantity."*

OT1 explained briefly that he used his *"experience"* from his training and from his real life to interpret the ICAO descriptors. OT2 put them in terms of percentage. However, he admitted that he *"didn"t actually count"* and that might have made him miss something. OT3, OT4 and OT5 all had hard time explaining how they interpret the ICAO descriptors. *"It"s very similar"* and *"it"s not black and white"* for OT3 while *"it"s very in details, very close to each other"* for OT4. OT5 sighed many times and looked uncomfortable when being ask to explain his interpretation. He said in the beginning that he counted before admitting later that he did not count exactly. He said that he looked *"at the overall, looked at other criteria also"*.

OT1 said that he did it *"equally"* which means both qualitatively and quantitatively. OT2 did not clearly state what kind of measure he took but he said that he compared the mistakes made by the candidates with his notes and put them in terms of percentage.

So, it could be concluded that he did it quantitatively which was the same as OT4 who said that he also did it quantitatively by taking notes every time the mistakes were made (*"I took notes when it happened the first time, the second time, the third time. For „almost never" I took it as happened only once in the entire interview.")* OT3 and OT5 considered the descriptors qualitatively. OT3 also made an interesting remark that *"everybody got his own mean for level four"* then, after listening to the speech samples he would think if *"they match with his own level four interpretation"*.

Table 4.188 shows the operational/untrained raters" (OU) interpretation of the ICAO scale descriptors.

Table 4.188: The operational/untrained raters" (OU) interpretation of the ICAO scale descriptors

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Ql.	Qt.	Ql.	Qt.	Ql.	Qt.	Ql.	Qt.	Ql.	Qt.	
43. & 44. Descriptor interpretation i.e. qualitatively (Ql.) or quantitatively (Qt.)	x		x		x		x		x		***OU1:*** *"„Almost never" sounds stronger than „rarely" because it has the word „never"." "„Almost never" is „never"" „almost" is there just in case. For example, a candidate may get blank for a while and it creates a pause. It has nothing to do with language. It's because he is thinking of something else. Then I put it as „never"."* ***OU1:*** *"Qualitatively."* ***OU2:*** *(Very long thinking so the researcher skipped to the next question)* ***OU2:*** *"Qualitatively." "I looked the errors they made but I didn"t actually count." "How frequent they made those mistakes." "How fluent they spoke. I had to see too if the question was too difficult for them or not." "That"s why I*

listened thoroughly once to see if and how often they spoke unevenly. Then I got the overview picture of that person. After that I looked at each criterion to see in details." (Reluctantly accepted) "Actually it"s about the feeling because I didn"t actually count." "I also think that getting each different level has different effect on them. If you get level six, you don"t have to be tested again in your entire life. If you get level five, you"ll be fine for a few years. I put this into my consideration when I award the scores."

OU3: *"My „rarely" is more than „almost never", more than in terms of number." "I didn"t count. I used my gut feeling to judge if this guy matches „frequently" or „usually"."*
OU3: *"Qualitatively." "It couldn"t be tally by numbers."*

OU4: *"I put level four as my standard and rate accordingly." "„Four" is acceptable." "How about two out of ten times for „only sometimes"?" "I didn"t actually count." "Then I change to 50% of the speech. If they make errors not more than half, I"d give them higher then level four." "I gave this guy „five" because ... um ..." (Long thinking) "I used my feeling." "I can"t differentiate between „rarely" and „only sometimes" but I let them pass if they made small errors. I took note if they made obvious errors. It wasn"t that specific that how many times was „rarely". They were „forgivable" and „unforgivable" errors. But I took notes and had a look later."*

OU4: *"Qualitatively." "I took notes but did not count in details in terms of numbers. To me, this kind of testing is like a conversation. You may make some mistakes in a conversation but if you can convey your message, I mean if*

	they can understand each other, I don"t care about the wordings." ***OU5:*** *"It"s very difficult to do." (Very hesitant) "„Rarely" and „almost never" are pretty close." "For example for „almost never" I cut „almost" off to be just „never". That means they don"t make any mistake. For „rarely" ... (sighed) ... if it interferes with understanding, it"d come down to level four. Level four is „sometimes".„Rarely" also means „sometimes" but „very sometimes"" "„Sometimes" may be three times but „rarely" is just only once." "Actually these adverbs are open for everybody to interpret." "For me, I interpret „almost never" as „never". That means there was absolutely no interference." "„Rarely" might happen just once." "I didn"t actually count because I didn"t take notes. I listened to the overall. I looked if it was comprehensible. If it was but there was some deviation, it might be „rarely". It might be his bad luck that he had to speak this word often. It might be his habit to speak like that."* ***OU5:*** *"Qualitatively. I didn"t use my feeling. It was sort of what I heard."*

In his opinion, OU1 considered *"„almost never" is „never""* because he thought that "*„almost" is there just in case"*. OU2 could not explain his interpretation to the researcher. It took so long that the researcher decided to skip to the next question. OU3 simply said that he used his *"gut feeling to judge if this guy matches „frequently" or „usually""*. OU4 also admitted that he used his *"feeling"* and he did not *"actually count"*. Similar to OU1, OU5 thought of *„almost never"* as *„never"*. OU5 joined OU4 in the manner of „not actually count" the mistakes made by the candidates because of his strategy of not taking notes at all.

All operational/untrained raters said that they considered the descriptors in terms of quality. OU2 interestingly accepted that he used some other things besides the ICAO descriptors in his score awarding. He thought that *"getting each different level has different effect"* on the candidates. That was described as *"If you get level six, you don"thave to be tested again in your entire life. If you get level five, you"ll befine for a few years. I put this into my consideration when I award the scores."*

Table 4.189 shows if the linguistic/trained raters (LT) consulted the ICAO descriptors before listening to the speech samples.

Table 4.189: The linguistic/trained raters" (LT) consultation with the ICAO descriptors before listening to the speech samples

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
46. ICAO descriptor consultation before listening to the speech samples	x			x	x		x		x		***LT1:*** *"A quick look before listening."* ***LT2:*** *"No, I remember them well because I read them frequently."* ***LT3:*** *"Rarely because I'm quite familiar with them."* ***LT4:*** *"I scanned it once before listening."* ***LT5:*** *"Sometimes."*

All linguistic/trained raters seemed to be familiar with the details of the ICAO descriptors since they did not spend much time consulting the details before listening to the speech samples. LT2 said that she did not do it at all because she read them frequently. LT1 and LT4 just did *"a quick look"* and *"scanned it once"*. LT5 spent *"sometimes"* to consult the details.

Table 4.190 shows if the linguistic/untrained raters (LU) consulted the ICAO descriptors before listening to the speech samples.

Table 4.190: The linguistic/untrained raters" (LU) consultation with the ICAO descriptors before listening to the speech samples

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
45. ICAO descriptor consultation before listening to the speech samples	x		x		x		x		x		***LU1:*** *"I skimmed through it once."* ***LU2:*** *"Yes, once."* ***LU3:*** *"I scanned once."* ***LU4:*** *"Once."* ***LU5:*** *"I scanned them once." "I read what „six" meant and I didn"t get it."*

In spite of lacking experience with the ICAO descriptors, all linguistic/untrained raters said that they consulted the descriptors just once. LU5 also added that he read what „six" meant and he *"didn"t get it"*.

Table 4.191 shows if the operational/trained raters (OT) consulted the ICAO descriptors before listening to the speech samples.

Table 4.191: The operational/trained raters" (OT) consultation with the ICAO descriptors before listening to the speech samples

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
45. ICAO descriptor consultation before listening to the speech samples	x		x		x		x		x		***OT1:*** *"I studied the whole document for an hour before start listening."* ***OT2:*** *"No, I didn"t." "Before I started the first one, I read through all over the criteria once." "I wanted to recall all the details before I started."* ***OT3:*** *"I looked at it „before". I looked at it „during"but not very often. And I definitely looked at it again „after"." "„Before" I just had a look, okay, just to remind myself. And then „during" if I wasn"t sure." "I scanned the descriptor just once."* ***OT4:*** *"Yes. Roughly. We were taught in the class to study this before rating."* ***OT5:*** *"Before listening? Once."*

OT1 spent most time consulting the details of the descriptors as *"an hour"* before starting to listen. The others (OT2, OT3 and OT5) did it once while OT4 did it *"roughly"*.

Table 4.192 shows if the operational/untrained raters (OU) consulted the ICAO descriptors before listening to the speech samples.

Table 4.192: The operational/untrained raters‟ (OU) consultation with the ICAO descriptors before listening to the speech samples

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
45. ICAO descriptor consultation before listening to the speech samples	x		x			x	x		x		***OU1:*** *"Not as much as it should have been."* ***OU2:*** *"Thoroughly once."* ***OU3:*** *"No, I didn"t readat all."* ***OU4:*** *"Quite thoroughly once because there are main ideas of each one."* ***OU5:*** *"Thoroughly once."*

OU1 did not clearly state how often he consulted the details but he admitted that he did it *"not as much as it should have been"*. OU2, OU4 and OU5 did it once while OU3 did not read them at all.

Table 4.193 shows if the linguistic/trained raters (LT) consulted the ICAO descriptors during listening to the speech samples.

Table 4.193: The linguistic/trained raters‟ (LT) consultation with the ICAO descriptors during listening to the speech samples

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
46. ICAO descriptor consultation	x			x	x			x	x		***LT1:*** *"Frequently during listening."*

during listening to the speech samples	***LT2:*** *"No, I didn"t, just listened."* ***LT3:*** *"Sometimes."* ***LT4:*** *"Not at all."* ***LT5:*** *"Sometimes."*

During listening to the speech samples LT2 and LT4 said that they did not consult the details of the descriptors while LT3 and LT5 said that they did it *"sometimes"*. LT1 was the only rater in this category who stated that she did it *"frequently"* during listening.

Table 4.194 shows if the linguistic/untrained raters (LU) consulted the ICAO descriptors during listening to the speech samples.

Table 4.194: The linguistic/untrained raters" (LU) consultation with the ICAO descriptors during listening to the speech samples

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
46. ICAO descriptor consultation during listening to the speech samples	x			x		x		x		x	***LU1:*** *"More often because when I really experienced the speech sample I wanted to know what level it should be."* ***LU2:*** *"No, not while listening."* ***LU3:*** *"No, I didn"t bok at it during listening."*

	LU4: *"Not at all. I just listened."*
	LU5: *"No."*

Four out of five linguistic/untrained raters (LU2, LU3, LU4 and LU5) said that they did not consult the details of the descriptors at all during listening to the speech samples. LU1 was the only rater who did it *"more often"* to see *"what level it should be"*.

Table 4.195 shows if the operational/trained raters (OT) consulted the ICAO descriptors during listening to the speech samples.

Table 4.195: The operational/trained raters" (OT) consultation with the ICAO descriptors during listening to the speech samples

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
46. ICAO descriptor consultation during listening to the speech samples	x		x		x			x	x		***OT1:*** *"If they are in between the levels, then I just consult the document again."* ***OT2:*** *"Often."* ***OT3:*** *"During, you know, listened to the tape, I had a look at it when he was doing something I think he was making a mistake ... just to check what it says in the paper." "„Sometimes" during."* ***OT4:*** *"No, not at all. Just took notes."*

	OT5: *"During listening? Almost ... well, I mean I listened to the mistakes being made and put them in there. So I constantly looked at the scales and looked at the speech. So frequently, I guess."*

Most of the operational/trained raters seemed to consult the details of the descriptors more often than the other groups. OT1 said that he did it *"if they are in between the levels"*. OT3 did it *"sometimes"* during listening to the speech samples *"to check what it says in the paper"*. OT5 *"frequently"* looked at the scales while OT2 did it *"often"*. OT4 was the sole rater in this batch who did not consult the details at all.

Table 4.196 shows if the operational/untrained raters (OU) consulted the ICAO descriptors during listening to the speech samples.

Table 4.196: The operational/untrained raters" (OU) consultation with the ICAO descriptors during listening to the speech samples

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
46. ICAO descriptor consultation during listening to the speech samples	x		x		x			x	x		***OU1:*** *"Quite often."* ***OU2:*** *"Once in a while."* ***OU3:*** *"I read them sometimes."* ***OU4:*** *"No I didn"tbecause I couldn"tseparate*

	my concentration."
	OU5: *"I kept checking while listening too."*

OU1 *"often"* and OT5 *"kept checking"* the details of the descriptors during listening to the speech samples while OU2 did it *"once in a while"* and OU3 did it *"sometimes"*. OU4 was the operational/untrained alone who did not consult the details at all because he could not separate his concentration i.e. he could not focus on more than one thing at a time.

Table 4.197 shows if the linguistic/trained raters (LT) consulted the ICAO descriptors after listening to the speech samples.

Table 4.197: The linguistic/trained raters" (LT) consultation with the ICAO descriptors after listening to the speech samples

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
47. ICAO descriptor consultation after listening to the speech samples	x		x		x		x		x		***LT1:*** *"Yes, always."* ***LT2:*** *"Yes, before I made the decision."* ***LT3:*** *"I frequently check it thoroughly even though I remember them."* ***LT4:*** *"Yes. I remembered them well but, I don"t know why, I still had to look at them."* ***LT5:*** *"Frequently."*

All linguistic/trained raters said that they consulted the details of the descriptors after listening to the speech samples before making their decision even though they remembered them (LT3 and LT4).

Table 4.198 shows if the linguistic/untrained raters (LU) consulted the ICAO descriptors after listening to the speech samples.

Table 4.198: The linguistic/untrained raters‟ (LU) consultation with the ICAO descriptors after listening to the speech samples

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
47. ICAO descriptor consultation after listening to the speech samples	x		x		x		x		x		***LU1:*** *“Most often because it was the time I had to give them the scores.”* ***LU2:*** *“Once again.”* ***LU3:*** *“I turned to it again after listening before awarding the scores.”* ***LU4:*** *“Always.”* ***LU5:*** *“Yes, once again. After listening I clearly got the picture of what „four‟meant. It might not be crystal clear but at least I got some ideas.”*

All linguistic/untrained raters said that they consulted the details of the descriptors at least once (LU2 and LU5) after listening to the speech samples before making their decision.

Table 4.199 shows if the operational/trained raters (OT) consulted the ICAO descriptors after listening to the speech samples.

Table 4.199: The operational/trained raters" (OT) consultation with the ICAO descriptors after listening to the speech samples

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
47. ICAO descriptor consultation after listening to the speech samples		x	x		x		x		x		***OT1:*** *"No. I didn"t go back and review the material."* ***OT2:*** *"Yes. Before I gave the score, I had to consult quite often."* ***OT3:*** *"Before giving the grade, I looked at it again, okay, what did he say overall and then go through, okay, pronunciation, it wasn"t level six. Was it level five? No. was is level four? I'm not sure. Was it level three? No, no, it"s better than level three. So „four'" "A whole lot „after"."* ***OT4:*** *"Yes, in more details but not thoroughly because they are too much."* ***OT5:*** *"After listening? Also frequently."*

Almost all operational/trained raters (OT2, OT3, OT4 and OT5) said that they consulted the details of the descriptors at least once (LU2 and LU5) after listening to the speech samples before making their decision. OT1 was the only rater in this group who said that he did not do that.

Table 4.200 shows if the operational/untrained raters (OU) consulted the ICAO descriptors after listening to the speech samples.

Table 4.200: The operational/untrained raters‟ (OU) consultation with the ICAO descriptors after listening to the speech samples

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
47. ICAO descriptor consultation after listening to the speech samples	x		x		x		x		x		***OU1:*** *"Before giving the final scores, I had another glance because I already read it during listening."* ***OU2:*** *"Frequently."* ***OU3:*** *"Frequently."* ***OU4:*** *"Once again. For example, if I'm gonna give the guy „five" I'd check again if „five" in pronunciation matched what I had in my mind. If it did, okay, „five" he"d get."* ***OU5:*** *"Once again to see if it was okay to award them such scores."*

OU1, OU4 and OU5 said that they consulted the details of the descriptors at least once after listening to the speech samples before making their decision. OU2 and OU3 said that they did it *"frequently"*.

Table 4.201 shows the linguistic/trained raters‟ (LT) opinion if every English native speaker must also be at ICAO Level 6.

Table 4.201: The linguistic/trained raters" (LT) opinion if every English native speaker must also be at ICAO Level 6

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
48. Every English native speaker must also be ICAO Level 6		x	x			x		x		x	***LT1:*** *"No. It depends on what level of education he has, what social status he has, what kind of life style he has." "If he"s in lower level, he may not get the appropriate vocabulary even if he is a native speaker of English."* ***LT2:*** *"Umm…I think so because the ability of level six is still below native speaker ability. Therefore native speaker ability is higher than ICAO level six."* ***LT3:*** *"Not necessarily because if they are backpackers and they don"t have accuracy even though they are natives. We have to use accuracy as a benchmark."* ***LT4:*** *"No, not necessarily because even some native speakers can make a lot of mistakes. ICAO uses a phrase that you must be intelligible to aeronautical community. It doesn"t mean that everybody has to be a native speaker of English. Therefore when some natives speak English, it"s possible that another native who lives in the other part of the world may not be able to understand them."* ***LT5:*** *"It"s hard to answer. Actually I"d like to answer „yes" but there was a native who was not rated as „six"." "My answer is „no" because there may be some factors during testing such as … I don"t know if you have heard this … an Australian pilot was rated by a Malaysian rater as „three". I feel that it shouldn"t be possible but there are some factors such as*

	natives who aren't well-educated. That may make some well-educated raters put the standard too high." "Raters may not understand some accents or some kinds of vocabulary that natives use. They may rate them as low as „five" but it shouldn"t be as low as „three" like this Malaysian rater did." "It depends on the educational level of both test-takers and raters."

When being asked if every English native speaker must also be at ICAO Level 6, four linguistic/trained raters said „no" with different reasons. It was *"depending on level of education he has"* for LT1, *"if they are backpackers and they don"t have accuracy even though they are natives"* for LT3, *"even some native speakers can make a lot of mistakes"* for LT4 and *"there was a native who was not rated as „six""* for LT5. The sole rater in this category who said „yes" was LT2 with the reason as *"the ability of level six is still below native speaker ability. Therefore native speaker ability is higher than ICAO level six".*

Table 4.202 shows the linguistic/untrained raters" (LU) opinion if every English native speaker must also be at ICAO Level 6

Table 4.202: The linguistic/untrained raters" (LU) opinion if every English native speaker must also be at ICAO Level 6

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
48. Every English native speaker must also be ICAO Level 6		x		x		x		x		x	***LU1:*** *"Not necessarily." "He may be level five." "Because it may depend on the flying experience. Even if he is a native, he may not be exactly at what stated in the ICAO criteria." "If that native is not a pilot, no, because he wouldn"t have the knowledge in this field."* ***LU2:*** *"No because it doesn"t mean that not all native speakers*

	of English can speak „good" English." "They may have different accents or they may not be good at speaking." "I feel that they may be „six" in pronunciation but may not be in grammar." "They may not be up to that level in some criteria." ***LU3:*** *"Not necessarily because he may get „six" in some criteria such as pronunciation." "It also depends on who performs the rating." "I may not understand a Scot. I may not be familiar with his pronunciation." "He may get a „six" in pronunciation and fluency but … not in comprehension. It's like we are Thai but we may not get full score in Thai. So do the English native speakers."* ***LU4:*** *"Definitely not, because some native speakers may not have the language proficiency in terms of basic grammar, complex structure to gain this level." "It's about a language use in a situation. They may have the intonation and pronunciation but when they operate in the real situation, it"sa language use in this specific context, they may not be able to use the language. They may not even know the concept so how could they use the language?"* ***LU5:*** *"I don"t think so because it doesn"t assess purely language proficiency. It also assesses aviation knowledge. Even very fluent guys may have problem with comprehension, vocabulary, something like this if they don"t have aviation knowledge."*

All linguistic/untrained raters" answers to the question if every English native speaker must also be at ICAO Level 6 were negative. LU1 thought that it was *"because it may depend on the flying experience"* and *"even if he is a native, he may not be exactly at what stated in the ICAO criteria".* It was because *"it doesn"t mean that not all native speakers of English can speak „good"*

English" for LU2. LU3 had another perspective of getting level 6 as *"it also depends on who performs the rating"*. LU4 was quite confident to say that *"definitely not, because some native speakers may not have the language proficiency in terms of basic grammar, complex structure to gain this level"* while it was because this kind of assessment *"doesn"t assess purely language proficiency" "it also assesses aviation knowledge"* for LU5.

Table 4.203 shows the operational/trained raters" (OT) opinion if every English native speaker must also be at ICAO Level 6.

Table 4.203: The operational/trained raters" (OT) opinion if every English native speaker must also be at ICAO Level 6

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
48. Every English native speaker must also be ICAO Level 6	x			x	x	x		x		x	***OT1:*** *"Ah…the ICAO level six is actually a lot easier than many English proficiency tests. So I believe that most natives going to be level six." "So, yes. But if you"re talking about …"* ***OT2:*** *"No because you can be a linguist or a native speaker … well… if you"ve been to a high school in the United States with me, (laughter) some people"s vocabulary aren"t as good as others or the structure. To be a native speaker or to be a linguist is totally different. You can speak the language…" "I don"t think so." "I don"t agree with that because you might not know the vocabulary. You might have troubles in some certain areas, especially the technical terms used in aviation."* ***OT3:*** *(Thinking) "No because I know some like (laughter) … let"s say, for example, like red necks, some people in the outback of Australia. They are native speakers but they can"t speak proper English. So it"snot a given(?) just because you"re*

a native, you get a level six. You"ve got to show you can do level six as well."
(On a second thought) "Well, maybe. Because everybody"s interpretation of a native speaker ... somebody who is ... very good English, got the accent, got the vocab. So if he gives him a level six, so ... probably native. You know he"sprobably born and raised ... he is pretty much native." "You work your way up to level six but you don"t automatic at the level six. Level six, you can only come down with level six."

OT4: *"No. Not necessarily. Because native speakers may come from many different countries, from Ireland, from Scotland and some of these accents may be very weird. Some people from international community may not understand them." "They may use slang or they may not be well educated so they use some terms which unintelligible to others who are non-native speakers of English. Not just pilots but those who are non-natives from other countries."*

OT5: *"No, I guess not. Well, I mean he could have ... some of his slang that affects the rating. You could use the language that sometimes is not understood by the interviewer or rater that he might not get a „six". I mean you could get „six" in fluency, „six" in comprehension, „six" in pronunciation, „six" in interactions. So I think ... not „six" ..." "Let"s say he has a southern or northern accent. If his accent affects his pronunciation and affects the meaning to the general English pronunciation then I guess it has something to do with the rating. He would not get a „six" for his pronunciation." "No because his accent may affect his pronunciation which affects overall rating."*

OT1 thought that every English native speaker must be at ICAO Level 6 because *"the ICAO level six is actually a lot easier than many English proficiency tests"*. The other operational/trained raters had different ideas. OT2 thought that native speakers *"might have troubles in some certain areas, especially the technical terms used in aviation"*. OT3 had two thoughts. On one hand, his first one was *„no"* because native speakers had to show they could do level six too. On the other hand, his second thought was *„maybe"* because of the *"everybody"s interpretation of a native speaker ... somebody who is ... very good English, got the accent, got the vocab"*. For OT4 it was *"not necessarily"* because *"native speakers may come from many different countries"* and *"they may use slang or they may not be well educated so they use some terms which unintelligible to others who are non-native speakers of English"*. OT5 had a similar view that some natives *"could have ... some of slang that affects the rating"*.

Table 4.204 shows the operational/untrained raters" (OU) opinion if every English native speaker must also be at ICAO Level 6.

Table 4.204: The operational/untrained raters" (OU) opinion if every English native speaker must also be at ICAO Level 6

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
48. Every English native speaker must also be ICAO Level 6		x		x		x		x		x	***OU1:*** *"Not necessarily because it"s about the discipline in using a language too." „This is ICAO, not ordinary English, in the part of radiotelephony. When they say „confirm", we have to answer „affirm"." "It"s the same in other part." "Even if ones graduate in linguistics, they may speak elaborately but it has nothing to do with „prœise"and „œoncise". It"s more than ICAO requirement and might mislead the meaning and objective of the ATC." "Natives may use incorrect standard phraseology."*

OU2: *"Not necessarily because you"re not just saying. You must make the others understand what you"re saying too. Otherwise there"ll be a communication breakdown. It doesn"t only depend on the speaker. It also depends on the listener. If the listener"s English is not good and you keep talking, you can"t get your listener to understand what you say."*

OU3: *"No. (Long thinking) Because aviation English has become global so native English speakers like ,,Speed bird" pilots may have difficulties when speaking with an Indian ATC." "I think there are many levels of English native speakers even among them." "I don"t feel comfortable to say that. It"s too conclusive. Level six has its own criteria that not all native speakers are able to match." "They may not pass the criterion of interactions."*

OU4: *"Not always because even some natives are unintelligible. They can"t communicate in words." "Language is a means to convey but before this, it"s a thought of that person. It"s how he can use language to convey his thought. It"s like a picture, a drawing. An architect uses his drawing to convey his thought without using language." "I have a friend who is an English native speaker but I hardly understand what he wants to convey." "For example, laypersons wouldn"t understand terms used in the field of education. Are they natives? Yes, they are but they use different kind of language." "Not all natives are good in grammar. They may use ungrammatical structured sentences but comprehensible. In this case, they wouldn"t get ,,six"according to the descriptors."*

OU5: *"No because some natives speak badly. They use incorrect grammar. They may not be well-educated."*

All operational/untrained raters had the same point of view about being an English native speaker and being at ICAO Level 6 that it was *"not necessarily"* for both of them to come together. OU1 thought that because *"it"s about the discipline in using a language too"* and, probably as a pilot, he additionally thought that *"natives may use incorrect standard phraseology"*. It was *"if the listener"s English is not good and you keep talking, you can"t get your listener to understand what you say"* for OU2. *"Level six has its own criteria that not all native speakers are able to match"* and *"They may not pass the criterion of interactions"* was OU3"s reason. OU4 and OU5 had similar standpoints as *"even some natives are unintelligible"* and *"some natives speak badly"*.

Table 4.205 shows the linguistic/trained raters" (LT) opinion if ICAO Level 6 is equivalent to an English native speaker.

Table 4.205: The linguistic/trained raters" (LT) opinion if ICAO Level 6 is equivalent to an English native speaker

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
49. Being at ICAO Level 6 is equivalent to being an English native speaker		x		x		x		x	x		***LT1:*** *"Level six allows some interferences from the first language so Thai students who study in high educational level or have exposure to native speakers." "Almost but not exactly." "Because those who are in level six should be able to organize their ideas in the same manner as native speakers do."* ***LT2:*** *"No because considering from the criteria, natives are a little better."* ***LT3:*** *"Some Thai students never study overseas but they can use correct and formal language without interference with the meaning. I can give them level six." "Level six doesn"t mean native speaker." "A Thai can be level six, he*

	doesn"t have to be a native because being a native means more than that. They are more natural." ***LT4:*** *"No because ICAO level six is not high. They can make a lot of mistakes. It"s allowed to make mistakes in every area."* ***LT5:*** *"If they non-native, they are equivalent." "Yes because ...if we read the level six descriptors, they are native-like." "I think so because those who make level six ... there"s nothing their interlocutors wouldn"t understand. They"re okay in every aspect." "I don"t really understand what they actually are." "ICAO level six descriptors are clearly defined as „native" or „native-like". It can be understood by native speakers." "Yes, I imply by myself that level six is „native" or „native-like"."*

Four linguistic/trained raters did not think that being at ICAO Level 6 was equivalent to being an English native speaker with various reasons. For LT1, it was *"almost but not exactly." "Because those who are in level six should be able to organize their ideas in the same manner as native speakers do."* LT2 thought that *"considering from the criteria, natives are a little better"* which was similar to LT3 whose thought was that *"a Thai can be level six, he doesn't have to be a native because being a native means more than that. They are more natural."* LT4 said *"no"* because *"ICAO level six is not high. They can make a lot of mistakes. It"s allowed to make mistakes in every area."* LT5 was the sole rater who said that *"they are equivalent"* because *"if we read the level six descriptors, they are native-like".* It is worth noting that she confessed later that *"Yes, I imply by myself that level six was „native" or „native-like"."*

Table 4.206 shows the linguistic/untrained raters" (LU) opinion if ICAO Level 6 is equivalent to an English native speaker.

Table 4.206: The linguistic/untrained raters‟ (LU) opinion if ICAO Level 6 is equivalent to an English native speaker

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
49. Being at ICAO Level 6 is equivalent to being an English native speaker	x			x		x		x		x	***LU1:*** *"If he is a Thai, even if he is good at grammar but his discourse …, he wouldn't get „six" because it also concerns cultural Sub-themes." "I'd give him „5.5"" "I consider being „six" as equivalent to an English native speaker."* ***LU2:*** *"Umm…not quite. For example he may get „six" in this criterion while a native may get only „five" because there may be some factors such as he may have background knowledge in the questions we ask so he would be able to answer fluently and confidently. At the same time, the native may not know much about the topic then he may not be so fluent." "As a matter of fact, I'd like to say „yes" because he fulfils the criteria. On the other hand, I'd like to say „no" because I'm not sure if the knowledge I have is enough to award him a „six". That makes me hesitate."* ***LU3:*** *"Not necessarily because those six criteria may not cover everything. Being native speakers may require more than that. Perhaps a Thai may get a „six" but he still can"t compare to a native speaker."* ***LU4:*** *"Um … it"s hard to say." "Level six, in terms of pronunciation, may still be influenced by their first language but understanding can be achieved successfully. So it doesn"t have to be a native speaker to achieve this level."*

	LU5: *"They may not be equivalent in terms of pronunciation, interactions, fluency, those sorts of things."*

LU1 admitted that she considered *"being „six" as equivalent to an English native speaker"* because *"it also concerns cultural Sub-themes"*. LU2 accepted that she was not sure if she had enough knowledge to award someone a „six". LU3 thought that it was *"not necessarily"* because *"those six criteria may not cover everything"* and *"being native speakers may require more than that"*. LU4 seemed to comply with the ICAO descriptors by saying that *"level six, in terms of pronunciation, may still be influenced by their first language but understanding can be achieved successfully. So it doesn"t have to be a native speaker to achieve this level"*. LU5 did not specify much in details. She just said that *"they may not be equivalent in terms of pronunciation, interactions, fluency, those sorts of things"*.

Table 4.207 shows the operational/trained raters" (OT) opinion if ICAO Level 6 is equivalent to an English native speaker.

Table 4.207: The operational/trained raters" (OT) opinion if ICAO Level 6 is equivalent to an English native speaker

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
49. Being at ICAO Level 6 is equivalent to being an English native speaker		x		x		x		x		x	***OT1:*** *"No." "Because it"s much easier."* ***OT2:*** *"No, because I think level six, let"s say pronunciation, you don"t have to be a native speaker. As long as you"re able to speak and others understand it, then you meet the criteria. You don"t have any mistakes in your structure. Then it"ll be okay to be level six." "You can be a level six but you might have problems with the accent." "If his accent is so strong that it"s hard to understand him. That*

could probably not make him a level six, even though he is a native speaker."

OT3: *"No. I'd say like, for example, Henry Kissinger. He speaks with a really heavy kind of Jewish kind of low voice but I'd say he"slevel six "œause he speaks really, really well but he doesn"tsound native. He doesn"tsound like a native speaker." (On a second thought) "Probably."*

OT4: *"No because of the same reasons as the previous question. Native speakers don"t have to be level six and level six don"thave to be natives." "Level six people may be able to use the language to communicate understandably in terms of grammar, fluency, vocabulary, choice of words. However, their pronunciation may be influenced by their first languages, by their original accents but it"s still correct."*

OT5: *"No. Is it the same question?" "No. Well, he could fulfill all the criteria for level six so that he gets a level six. It doesn"t mean that he has to be a native speaker. He doesn"t have to be an American. He doesn"t have to be English. He could be French. If he uses pronunciation as what is required for level six, then he gets a level six. It doesn"tmean that he has to be a native speaker." "Or even for a Thai, if he"sfluent enough for „six",if he"s… ah … his vocabulary is for „six",he might study abroad, he might be fluent, he could get a „six".He doesn"thave to be a native speaker." "He gets level six because he fulfills all criteria for level six."*

All operational/trained raters said „no" to the question if being at ICAO Level 6 was equivalent to being an English native speaker with diverse reasons. OT1 thought that because *"it"s much easier"*. OT2"s idea was that *"as long as you"re able to speak and others understand it, then you meet the criteria"*. OT3 was uncertain as he initially said *„no"* then changed to *„probably"* later. OT4 confirmed his opinion by saying that *"native speakers don"t have to be level six and level six don"t have to be natives"* because *"level six people may be able to use the language to communicate understandably in terms of grammar, fluency, vocabulary, choice of words. However, their pronunciation may be influenced by their first languages, by their original accents but it"s still correct"*. OT5 showed his understanding of the ICAO criteria by saying that *"he doesn"t have to be a native speaker." "He gets level six because he fulfills all criteria for level six."*

Table 4.208 shows the operational/untrained raters" (OU) opinion if ICAO Level 6 is equivalent to an English native speaker.

Table 4.208: The operational/untrained raters" (OU) opinion if ICAO Level 6 is equivalent to an English native speaker

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
49. Being at ICAO Level 6 is equivalent to being an English native speaker		x		x		x		x		x	***OU1:*** *"No because they can get only the ability to communicate comprehensibly, easy, precisely and concisely concerning aviation only."* ***OU2:*** *"No. Maybe close to but not equivalent."* ***OU3:*** *"No because they"re different issues. Native speakers … it"s a quality that ICAO says that pilots should have. Native speakers may not have it. People with level six don"t have to be native speakers." "Native speakers are like mother-tongue. Level six is like operational criteria."*

	"Level six is like a hurdle for pilots to jump across. It"s like a measure stick. It doesn"t mean that you"re a native speaker if you can cross it."
	OU4: *"Not always because he may not get „six" in every criterion but his overall performance is „six". If he gets all „six", he"s close to a native. But as I said, we still think in our own language. We may have strategies to translate quickly and use language correctly and superbly but as far as we don"t think in that language, there"ll be something to show that we"re not native, just close to be."*
	OU5: *"No because it"s up to the criteria. Even though level six seems to be perfect in every way but, for example, it doesn"t say that pronunciation and accent are limited to be native speaker"s accent. They may be influenced by their first language but it almost never interferes. As long as it does not interfere with the understanding, they get level six while native speakers may not have any accent. So it couldn"t be said that level six is equivalent to a native speaker because level six may have local foreign accent." "The criteria do not say that." "Level six have no problem using the language in their job but it may not be as elaborate and profound as native speakers." "Being natives has more intense ability in using the language than being level six. Those natives who are well educated are better than level six."*

All operational/untrained raters agreed that being at ICAO Level 6 was not equivalent to a native speaker of English with varied reasons. For OU1, being at ICAO Level 6 just meant that *"they can get only the ability to communicate comprehensibly, easy, precisely and concisely concerning aviation only"*. OU2 did not explain his reason, just said that it might be *"close to but not*

equivalent". "Level six is like operational criteria" for OU3 and *"Level six is like a hurdle for pilots to jump across. It"s like a measure stick. It doesn"t mean that you"re a native speaker if you can cross it."* It was *"not always"* for OU4. His perspective was that *"if he gets all „six", he"s close to a native"*. However, *"we still think in our own language. We may have strategies to translate quickly and use language correctly and superbly but as far as we don"t think in that language, there"ll be something to show that we"re not native, just close to be."* OU5 said *"no"* because *"it"s up to the criteria"* and *"Level six have no problem using the language in their job but it may not be as elaborate and profound as native speakers."*

Table 4.209 shows if the linguistic/trained raters (LT) were aware that Level 4 was the cut-off score.

Table 4.209: The linguistic/trained raters" (LT) awareness of Level 4 as the cut-off score

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
50. Awareness of Level 4 as the cut-off score	x		x		x		x		x		***LT1:*** *"Yes."* ***LT2:*** *"Yes."* ***LT3:*** *"Yes." (Laughter)* ***LT4:*** *"Yes."* ***LT5:*** *"Yes."*

All linguistic/trained raters knew that the „cut-off" score for the ICAO assessment was level 4.

Table 4.210 shows if the linguistic/untrained raters (LU) were aware that Level 4 was the cut-off score.

Table 4.210: The linguistic/untrained raters" (LU) awareness of Level 4 as the cut-off score

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
50. Awareness of Level 4 as the cut-off score		x		x		x	x			x	***LU1:*** *"No, I don"t."* ***LU2:*** *"No, I don"t"* ***LU3:*** *"No."* ***LU4:*** *"Yes. It"s stated in the ICAO scales."* ***LU5:*** *"No, I don"t."*

Almost all linguistic/untrained raters said that they did not realize that the „cut-off" score for the ICAO assessment was level 4 except LU4 who mentioned that it was stated in the ICAO scales.

Table 4.211 shows if the operational/trained raters (OT) were aware that Level 4 was the cut-off score.

Table 4.211: The operational/trained raters" (OT) awareness of Level 4 as the cut-off score

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
50. Awareness of Level 4 as the cut-off score	x		x		x		x		x		***OT1:*** *"Yes."* ***OT2:*** *"Yes."* ***OT3:*** *"Yes."* ***OT4:*** *"Yes."* ***OT5:*** *"Yes."*

All operational/trained raters realized that the „cut-off" score for the ICAO assessment was level 4.

Table 4.212 shows if the operational/untrained raters (OU) were aware that Level 4 was the cut-off score.

Table 4.212: The operational/untrained raters" (OU) awareness of Level 4 as the cut-off score

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
50. Awareness of Level 4 as the cut-off score	x		x			x	x		x		***OU1:*** *"Yes."* ***OU2:*** *"Yes."* ***OU3:*** *"No, I didn"t"*

OU4: *"Yes."*

OU5: *"Yes."*

Almost all operational/untrained raters said that they knew that the „cut-off" score for the ICAO assessment was level 4 except OU3 who was unaware of that score.

Table 4.213 shows if the linguistic/trained raters (LT) considered the consequences as „pass" or „fail" in their ratings.

Table 4.213: The linguistic/trained raters" (LT) consideration of the consequences as „pass" or „fail" in their ratings

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
51. Consideration of the candidates" consequences as "pass" or "fail" in ratings		x		x	x		x		x		***LT1:*** *"No." "Because we rate as what are stated." "The candidate may do better next time or the third time after they"re more familiar with the test environment and they don"t get excited or nervous."* ***LT2:*** *(Laughter) "No because otherwise we"d be sympathetic listeners." "I don"t wantto be biased."* ***LT3:*** *"I do consider because they have to retest. What I can do to help them is to give the comments that I have for them. What I jotted down can help them improve their English."* ***LT4:*** *"It must be considered." "My consideration is that if you fail, you must improve your proficiency. It doesn"t mean that you"ll lose your job if you fail. It*

	means that you"ll have to improve your English if you fail. My consideration is the language, not the career." ***LT5:*** *"Yes because I don"t want to see them fail. It"s because in real life it means their career if they can carry on their job or not." "It"s likely that I may round up the scores."*

Two linguistic/trained raters (LT1 and LT2) said that they did not consider the consequences of the candidate as being „pass" or „fail" in their ratings while the other three (LT3, LT4 and LT5) stated that they thought about it with different perspectives. On one hand, LT3 and L4 considered it in terms of the candidates" language – how could they improve their English? (*"What I jotted down can help them improve their English"* – LT3 and *"It means that you"llhave to improve your English if you fail"* – LT4) On the other hand, LT5 regarded it in terms of the candidates" career as *"I don"twant to see them fail. It"s because in real life it means their career if they can carry on their job or not."* Moreover, she even said that it was likely that she might *"round up the scores."*

Table 4.214 shows if the linguistic/untrained raters (LU) considered the consequences as „pass" or „fail" in their ratings.

Table 4.214: The linguistic/untrained raters" (LU) consideration of the consequences as „pass" or „fail" in their ratings

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
51. Consideration of the candidates" consequences as "pass" or "fail" in ratings		x		x		x	x		x		***LU1:*** *"It is a chance." "Umm…as a rater, I would insist on the given scores." "But being me, if I know (hesitant) I wouldn"tbecause I do my job as a rater so I shouldn"t do that."*

	LU2: *"No because this is what you are." "I have the feeling that the first score I gave is the correct one." "I don"t care if he passes or fails."* ***LU3:*** *"No because it would be worse if he has problems and he can"t communicate." "I concern more about his professional responsibility."* ***LU4:*** *"Yes. That"s right." "I don"t want to see him fail just because of his language ability only even though he has experience to operate his job and he can communicate by using the standard phraseology."* ***LU5:*** *"As I said, I wouldn"t because I still don"t know if he"s really good or not." "I weighed the score from just three criteria – fluency, interactions and pronunciation. Those are all three criteria I can assess as far as my background allows. I don"t care much about vocabulary and comprehension."*

Three linguistic/untrained raters (LU1, LU2 and LU3) said that they would not consider the consequences of the candidate as being „pass" or „fail" in their ratings. However, LU1 was quite hesitant before giving her final statement as *"I wouldn"t because I do my job as a rater so I shouldn"t do that."* LU2 insisted that she would not *"care if he passes or fails"* while LU3 seemed to be concerned about the candidates" career but in a different aspect from LT5. She determined that *"it would be worse if* (the candidate) *has problems* (during his line duty) *and he can"t communicate"* so she was concerned *"more about* (the candidate"s) *professional responsibility".* LU4 and LU5 stated that they would consider the candidate"s consequences but with dissimilar reasons. LU4 did not want to see any candidate fail *"just because of his language ability only even though he has experience to operate his job and he can communicate by using the standard phraseology"* while LU5 accepted that he considered that because he was uncertain if the

candidate was *"really good or not"*. He also interestingly admitted that he *"weighted the score from just three criteria – fluency, interactions and pronunciation"* since his background (as an English teacher) did not allow him to assess other criteria. Thus, he did not care *"much about vocabulary and comprehension"*.

Table 4.215 shows if the operational/trained raters (OT) considered the consequences as „pass" or „fail" in their ratings.

Table 4.215: The operational/trained raters" (OT) consideration of the consequences as „pass" or „fail" in their ratings

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
51. Consideration of the candidates" consequences as "pass" or "fail" in ratings		x	x		x		x		x		***OT1:*** *"I try not to." "Because if I had given a candidate level four and then, for some reasons, we were audited then there"d be more of the problem so I try just basically do the rating the way I think I should have been."* ***OT2:*** *"That was in my mind. So, yes, I would consider that." "Who would not consider that?" "I tried to look at it as my own airline as well." "I was also thinking about what it would be to push them up." "I think I wouldn"t bebiased. If the guy passed, he passed."* ***OT3:*** *(Sighed) "I did consider it. Because you don"t wanna be, you know, the man with the sword saying whether you can fly or not fly." (A pause) "I mean I did consider it but did it affect my scores? No."*

	OT4: *"Yes because it"s crucial to their career. It"s the matter of losing or retaining their jobs. They have their families to take care."* ***OT5:*** *"Well, I mean you know that ... a failing score would affect his somewhat future, his life but as a rater you would have to cut that off. I mean don"t think about that, I guess. You don"t even know who you"re rating. You just listen to the number being said, do your job and then I"d say ... even in your heart you might have some sympathy but ... generally no, I guess." "I would say deeply there is some room for sympathy." "„Don"t" would be too harsh, I guess." "I would have the same sympathy for everyone." "I mean I would consider but it wouldn"t affect my ratings."*

All operational/trained showed some degrees of reluctance or uneasiness when confronting with this question. OT1 accepted he tried not to consider that because he was concerned about the consequences if he would be audited later. OT2 admitted that the consideration was on his mind. He even challenged that *"who would not consider that?"* before concluding eventually that he would not be biased. OT3 sighed before acknowledging that he *"did consider"* still it did not affect his scores. OT4 seemed to more thoughtful in terms of the candidates" career *"because it"s crucial to their career. It"s the matter of losing or retaining their jobs. They have their families to take care".* OT5 had the same idea as OT2 that he *"would consider"* but *"it wouldn"t affect"* his ratings.

Table 4.216 shows if the operational/untrained raters (OU) considered the consequences as „pass" or „fail" in their ratings.

Table 4.216: The operational/untrained raters" (OU) consideration of the consequences as „pass" or „fail" in their ratings

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
51. Consideration of the candidates" consequences as "pass" or "fail" in ratings	x			x	x			x	x	x	***OU1:*** *"Yes because otherwise the aviation industry will collapse." "I don"t want to fail them." "How could we explain why pilots who have been flying for ten years without any accident or incident should fail in using English?" "They"ve made many takeoffs and landings." "I believe in their previous experience." "If they fail, it"ll have effect to the company too."* ***OU2:*** *"No because if you pass, you pass. If you fail, you fail." "But I won"t give level two or one, just level three."* ***OU3:*** *"I didn"t consider it from the beginning because I didn"t know… well, yes, I do now. I sympathize with them if they fail. In a real rating I would feel bad if they have to re-test again." "It can be considered in two ways. Consider that it"s the opportunity for their self-improvement."* ***OU4:*** *"I do." "Let me answer again. I know that „four" is pass, „three" or lower is fail. Do I consider this in my rating? Um … I don"t." "Because I awarded the scores according to their abilities." "In my mind, I thought that „four" had to re-test in three years." (Very hesitant) "I can fail them if I feel they deserve that." "I think it"s fair for everybody." "If we let everyone pass, we wouldn"t see where the weak points are." "Let"s say I don"t consider the consequences."*

	OU5: *"No because ... it"s raters" duty to assess. If all raters consider this before awarding the scores, they"d give „fair" to every candidate." "Yes, I might consider but I still give them that score as it is."*

OU1 seemed to concern much about non-linguistic aspects in his ratings. He admitted that he considered the candidates" consequences because *"otherwise the aviation industry will collapse"*. Moreover, he even raised a question - *"how could we explain why pilots who have been flying for ten years without any accident or incident should fail in using English?"* OU1 insisted that he believed *"in their previous experience."* OU2 had an opposite idea. He said that he would not consider the consequences *"because if you pass, you pass. If you fail, you fail"*. However, he added without any more explanation that he would not *"give level two or one, just level three."* OU3 who did not know from the beginning that the cut-off score was level 4 also accepted that he *"would feel bad if they have to re-test again"*. Even so, he still looked at the positive side by considering *"that it"s the opportunity for their self-improvement"*. OU4 was quite hesitant to answer this question. He said „yes" in the beginning before changing to „no" afterwards because he *"awarded the scores according to their abilities"* and *"it"sfair for everybody"*. OU5 also answered *"no"* in the beginning then changed to *"yes"* in the end. However, he added that he would *"still give them that score as it is"*.

Table 4.217 shows if the linguistic/trained raters (LT) considered any personal relationship with the candidates.

Table 4.217: The linguistic/trained raters" (LT) consideration of any personal relationship with the candidates

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
52. Consideration of		x		x		x		x	x		***LT1:*** *"No." "Because there are proper procedures and concrete evidence that can justify why I rate them as*

any personal relationship with the candidates	*such." "Not even if they are my superior."* ***LT2:*** *"No because every time I rate I consider that as my standard."* ***LT3:*** *"No because everything has evidence. If something happens and they trace back to see who does that rating, it wouldn"t benice and it"ll bemy fault."* ***LT4:*** *"No because we rate them in terms of language. It can be improved. If we help them, they wouldn"tget any improvement." "The rating has evidence. It can be matched with the scales." "It"s sort of ethic." "Actually we"rehelping them."* ***LT5:*** *"Yes." "If it can be rounded up, I'll do it but it must also be within an acceptable limit."*

Almost all linguistic/trained raters (LT1, LT2, LT3 and LT4) said that they would not consider changing the scores they already gave the candidates because they had any kind of relationship with them. They stated various reasons for their same answers. It was *"because there are proper procedures and concrete evidence that can justify why I rate them as such"* for LT1 while it was *"because every time I rate I consider that as my standard"* for LT2. LT3"s concern was the aftermath of her ratings *"because everything has evidence. If something happens and they trace back to see who does that rating, it wouldn"t be nice and it"ll be my fault"*. LT4 had a similar idea that *"the rating has evidence"* and *"it can be matched with the scales"*. LT5 was the sole rater in this group who admitted that she would consider changing the scores by rounding it up under the condition that *"it must also be within an acceptable limit."*

Table 4.218 shows if the linguistic/untrained raters (LU) considered any personal relationship with the candidates.

Table 4.218: The linguistic/untrained raters" (LU) consideration of any personal relationship with the candidates

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
52. Consideration of any personal relationship with the candidates	x			x		x		x		x	***LU1:*** *"It's hard to answer." "I would because it concerns personal relationship."* ***LU2:*** *"No because I look at this job as if „everybody"s life is in your hands". That means if you fail, you should consider improving yourself in order to give confidence to people who use your service. They must be totally confident in you."* ***LU3:*** *"No because of the same reason – their professional responsibilities."* ***LU4:*** *"No because I rate according to the scales and rubric."* ***LU5:*** *"This concerns teacher"s ethics that I have. So I wouldn"t."*

Only one linguistic/untrained rater (LU1) accepted that she would consider changing the scores because *"it concerns personal relationship"* while the other four said „no" because LU2 looked at this job *"as if „everybody"s life is in your hands""* which is similar to LU3 who looked at *"their professional responsibilities"* and LU4 rated *"according to the scales and rubric"*. LU5"s concern was the *"teacher"s ethics"* that he had.

Table 4.219 shows if the operational/trained raters (OT) considered any personal relationship with the candidates.

Table 4.219: The operational/trained raters" (OT) consideration of any personal relationship with the candidates

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
52. Consideration of any personal relationship with the candidates	-	-		x		x	x			x	***OT1:*** *"I would like to believe that I would not be influenced by that. However, I really do not know if I could do it. I honestly don"t know." "It"d be difficult. I don"t know the answer for that. To be truthfully said."* ***OT2:*** *"No because I think I took it all to myself ... if I was to help this guy I would make him better."* ***OT3:*** *(Sighed and paused) "I know the guys very well." (Long pause) "No." "Everybody"s got to have a certain standard. If he meets it, then he meets it. If he doesn"t meet it, he doesn"t meet it. Am I sorry? Yes. Would I like to help him? Yes. But I got my own standard and level four is level four." (Long pause) "That"s why I prefer not to be in this situation." (Laughter)* ***OT4:*** *"There are two cases. If I know in advance that it"s this guy, I wouldn"t give him „three". I"ll try to look at his strong points and try to give him „four"." "If I know later, I"d change the score." "The reason is as I said. Because I know him, I know his job and I know that he can do it." "The first reason is because I know him and because I know him so I know that he can handle his job. He is able to do his flight duties. He"s been doing it for some time."*

	OT5: *"No." "Well, I mean it would affect me. Every speech sample has its proof for the rating that he gets. If I overrate, it affects my judgment. If I underrate, it affects my judgment. So I'd rather be up to the standard level." "I won't change the scores."*

Three operational/trained raters (OT1, OT3 and OT4) seemed to be quite uncomfortable to answer this question. OT1 honestly admitted that he did not know the answer for this question. OT3 sighed and paused many times before answering „no" but he preferred *"not to be in this situation"*. OT4 accepted that he would change the scores because he knew the candidate and he knew that the candidate could do it. OT4 thought that the candidate could *"handle his job"* because *"he's been doing it for some time."* OT2 and OT5 said that they would not change the scores. OT5 gave the reason that it was because he would *"rather be up to the standard level"* because *"every speech sample has its proof for the rating that he gets"*.

Table 4.220 shows if the operational/untrained raters (OU) considered any personal relationship with the candidates.

Table 4.220: The operational/untrained raters' (OU) consideration of any personal relationship with the candidates

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
52. Consideration of any personal relationship with the candidates		x		x		x		x		x	***OU1:*** *"No. there's nothing to do with relationship. I'm already concerned about other factors."* ***OU2:*** *"No because it's not right."* ***OU3:*** *"No because of fairness." (After pondering) "If*

	they are my superior and I am anonymous, it is what it is. If I'm not anonymous, I'm not sure." (Hesitant) "I think it has effect on my rating." ***OU4:*** *"In my mind, I wanna help them. I guarantee that if we are friends or relatives, we want to help for sure. But this kind of testing there are evidence. We have speech samples. We have scoring sheets. Most of all we have comments why we award them with such scores. If we help them, we"ve got to change all these evidence." "We can"t change them because the fact is in the speech sample. Otherwise the rater himself is not up to standard." "I'll help them in some other ways. If they fail, I'll teach them." "I wouldn"tchange the scores." "If they are my superior, (Very unpleasant) I'd like to say that I'd refuse to answer this question." "I wouldn"t change because I"d think I don't know who they are. I'll try to maintain the standard of raters." "It'd be better if we pretend that we don"tknow who they are even though we do." "We should help them in some other ways." "If you ask if I'm afraid, yes, I am." "This is my personal answer because this is a norm in Thai society."* ***OU5:*** *"No because of the same answer as above. As a rater you have to do your duties." "Therefore raters shouldn"t bethe guys in the organization."*

All operational/untrained raters said that they would not consider changing the scores. OU1 was *"already concerned about other factors"*. OU2 and OU3 simply said that because *"it"snot right"* (OU2) and *"because of fairness"* (OU3). However, OU3 also confessed that *"if they are my superior and I am anonymous, it is what it is. If I'm not anonymous, I'm not sure"*. He was quite

hesitant before saying further that *"I think it has effect on my rating."* OU4 answered that he would not change the scores because *"I"d think I don"tknow who they are. I'll try to maintain the standard of raters."* OU4 added that *"it"dbe better if we pretend that we don"tknow who they are even though we do"* and *"we should help them in some other ways"*. However, he acknowledged *"if you ask if I'm afraid, yes, I am." "This is my personal answer because this is a norm in Thai society."* OU5 was determined to say „no" but he also said that *"raters shouldn"t bethe guys in the organization."*

Table 4.221 shows if the linguistic/trained raters (LT) were aware that the overall score must be the lowest score in a criterion.

Table 4.221: The linguistic/trained raters" (LT) awareness of the overall score as the lowest score among all six criteria

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
52. Awareness of the overall score as the lowest among all six criteria	x		x		x		x		x		***LT1:*** *"Yes."* ***LT2:*** *"Yes, I know."* ***LT3:*** *"Yes."* ***LT4:*** *"Yes."* ***LT5:*** *"Yes."*

All linguistic/trained raters accepted that they realized that the lowest scores among all six criteria were required by ICAO to be the overall for the candidates.

Table 4.222 shows if the linguistic/untrained raters (LU) were aware that the overall score must be the lowest score in a criterion.

Table 4.222: The linguistic/untrained raters" (LU) awareness of the overall score as the lowest score among all six criteria

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
52. Awareness of the overall score as the lowest among all six criteria		x		x		x		x		x	***LU1:*** *"No."* ***LU2:*** *"No, I don"t"* ***LU3:*** *"No."* ***LU4:*** *"No, I didn"t."* ***LU5:*** *"No."*

None of the linguistic/untrained raters perceived that ICAO requires the overall score to be based on the lowest score among all six criteria.

Table 4.223 shows if the operational/trained raters (OT) were aware that the overall score must be the lowest score in a criterion.

Table 4.223: The operational/trained raters' (OT) awareness of the overall score as the lowest score among all six criteria

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
52. Awareness of the overall score as the lowest among all six criteria	x		x		x		x		x		***OT1:*** *"Yes."* ***OT2:*** *"Yes."* ***OT3:*** *"Yes."* ***OT4:*** *"Yes."* ***OT5:*** *"Yes."*

Every operational/trained rater admitted that he knew the overall scores that the candidates would be awarded were based on the lowest among all six criteria.

Table 4.224 shows if the operational/untrained raters (OU) were aware that the overall score must be the lowest score in a criterion.

Table 4.224: The operational/untrained raters' (OU) awareness of the overall score as the lowest score among all six criteria

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
52. Awareness of the overall score as the lowest among all six criteria		x		x		x		x		x	***OU1:*** *"No. Really?"* ***OU2:*** *"No."*

OU3: *"No."*

OU4: *"No." "Oh! This is not fair."*

OU5: *"No."*

The whole lot of the operational/untrained raters said that they were not aware that the lowest scores among all six criteria would be the candidates‟ overall scores as required by ICAO. Some (OU1) was surprised and some even commented that it was unfair (OU4).

Table 4.225 shows if the linguistic/trained raters (LT) would consider changing the score after knowing that the overall score was based on the lowest score among all six criteria.

Table 4.225: The linguistic/trained raters‟ (LT) consideration to change the score after knowing that the overall score was based on the lowest score among all six criteria

Sub-themes	Rater LT1		Rater LT2		Rater LT3		Rater LT4		Rater LT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
53. Consideration of score change after knowing that the overall score is based on the	x			x		x		x	x		***LT1:*** *"As I said, for example, this guy who got „four" in most criteria, I would give him „four" instead of „three" if I knew nothing about the ICAO requirement that the overall score must be the lowest score from any criterion."* ***LT2:*** *"No, I won"tchange for this guy." "If he gets „four"*

lowest score among all six criteria	*in five criteria and „three" in just one criterion, then I have to consider if his „three" is very low „three" or not. I have to see if that interferes with the overall language or not." "There's a chance that I may change the score but it depends." "I mostly look at the first three criteria as main points. Another thing which is also important is comprehension."* ***LT3:*** *"We have to consider the objectives of this rating." "No, I wouldn't. As I told you, I equally put the importance to all six criteria but the first four will be shown first. If he makes a lot of mistakes, he wouldn't make it." "I wouldn't change." "I agree with ICAO."* ***LT4:*** *"No because each criterion was already carefully considered."* ***LT5:*** *"Yes, I would." "My consideration depends on each individual situation." "If they get higher scores in most criteria, I'd award him that higher score. For example, if he gets „five" in four criteria and „four" in the other two, I'd give him „five" instead of „four." But if he gets „four" in four criteria and „five" in the other two, I'd give him „four.""*

The linguistic/trained raters' opinions were diverse when being asked if they would change their scores in case that they knew nothing about the ICAO requirement of the overall scores to be based on the lowest scores among all six criteria. Two raters in this group (LT1 and LT5) said that they would change their scores. LT1 stated that she would give „four" to the guy who got „four" in most criteria and „three" in just one criterion while LT5 said in a similar way that she would award a candidate a higher score if he received

higher scores in most criteria. LT2 did not clearly say if she would change the scores or not. She said that she had to consider if that lower score was „how low" and she had to see *"if that interferes with the overall language or not"*. She added that she looked *"at the first three criteria as main points"* and *"another thing which is also important"* to her was *"comprehension"*. The other two (LT3 and LT4) had an opposing point of view. They said that they would not change their scores because they *"equally put the importance to all six criteria"* (LT3) and *"each criterion was already carefully considered"* (LT4).

Table 4.226 shows if the linguistic/untrained raters (LU) would consider changing the score after knowing that the overall score was based on the lowest score among all six criteria.

Table 4.226: The linguistic/untrained raters' (LU) consideration to change the score after knowing that the overall score was based on the lowest score among all six criteria

Sub-themes	Rater LU1		Rater LU2		Rater LU3		Rater LU4		Rater LU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
53. Consideration of score change after knowing that the overall score is based on the lowest score among all six criteria	x		x		x		x		x		***LU1:*** *"But it also depends on the score. It shouldn't be too bad. For example, if it's „three" not „two", I'd raise it to „four".* ***LU2:*** *"Yes, I may change." "I would change the overall to what I have in mind because it is overall. For example, if I think he should be „four", I'll change the overall score to „four" no matter what the other criteria are."* ***LU3:*** *"No, I wouldn't." (Hesitant) "I think I would change to lower scores." "I'd change this guy's score to „four" because he did well in that part. I think the first part is more important than the interview part. It*

	concerns his job." "But if he did badly in the first part, that"sthe end." "This guy could communicate in his flight duties but he had problems using English in his everyday life." "I weigh more on the first part. So if he does well in that part, I tend to change the score based on that part." ***LU4:*** *"Yes because I don"tthink he should fail because of his language ability only." "He is able to complete his task in the real situation."* ***LU5:*** *"Yes." "I would change the score for this guy from „thræ" to „four" to let him pass." "Because, as I told you, this guy could intelligibly communicate. I looked at the overall as the main consideration. I looked at mutual understanding if it can fulfill the objectives. So if I know that „three" means „fail", I would … as a teacher, this guy shouldn"t fail." "I put my overall score as the prime judgment. If the score in any criterion does not concur with it, I would change that score to correspond with my overall." "But I wouldn"tchange the score for this guy because, as a layperson, I look at the pilot profession that the core of this English for specific purpose use is safety and security. Therefore comprehension is considered crucial. If you can"tcommunicate, it"s the end. In this case, you can"t communicate, so you"dbetter fail." "In summary, I would change, just depending on the criteria I consider important."*

All linguistic/untrained raters admitted that they would change their scores. *"But it also depends on the score. It shouldn"t be too bad"* was LU1"s idea. LU2 said that she *"would change the overall to what I have in mind because it is overall"*. Furthermore, she added that *"if I think he should be „four", I"ll change the overall score to „four" no matter what the other criteria are"*. LU3 was

hesitant. In the beginning she said that she would not change, then she amended her answer that she *"would change to lower scores"*. LU3"s idea was similar to LU4"s in that both of them thought of the candidates" flight duties as the prime concern of their decisions. LU3"s consideration that *"this guy could communicate in his flight duties but he had problems using English in his everyday life"* and LU4"s statements that *"I don"t think he should fail because of his language ability only"* and *"he is able to complete his task in the real situation"* evidently showed their concerns. LU5 insisted that she put her overall scores as her prime judgment. *"If the score in any criterion does not concur with it, I would change that score to correspond with my overall"* exhibited her determination. She looked *"at the pilot profession that the core of this English for specific purpose use is safety and security. Therefore, comprehension is considered crucial"*. She summarized that she would change her scores *"just depending on the criteria"* that she considered *"important"*.

Table 4.227 shows if the operational/trained raters (OT) would consider changing the score after knowing that the overall score was based on the lowest score among all six criteria.

Table 4.227: The operational/trained raters" (OT) consideration to change the score after knowing that the overall score was based on the lowest score among all six criteria

Sub-themes	Rater OT1		Rater OT2		Rater OT3		Rater OT4		Rater OT5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
53. Consideration of score change after knowing that the overall score is based on the	-	-	x		x		x			x	***OT1:*** *(Very hesitant) "I don"tknow." "I think I'd have to listen to it, maybe two or three more times. I'll really, really try to give him four on that. I think I'll need more time to listen to the tape until I"mabsolutely sure."* ***OT2:*** *"Yes. I would give him a mean average. It could be 4.8" "I don"treally like the way that you got all „five"but*

lowest score among all six criteria	*you got one „four" you got a „four". The way ICAO does."* ***OT3:*** *"Let"sput it this way, if it was „four", „four", „four", „four", „four", „three", I'd consider give him „four". But if it was „fours", „five", it"d still „four."" (Sighed) "It depends on my overall impression. It'll be like I said taking it as a whole." "If he gets all „four" just one „three", I'd bump up rather from „three" to „four" rather than from „four"to „five". I mean that if he was overall „four", „four", „four", „four" with one „three", I'll be more cline to bump that to „four", just because I know that he should be „four" and I know that „four" is minimum requirement." "If the guy gets all „four"but one „three", I have to listen to the tape again, just to see whether I could find anything that can bump to „four"but would I change it just for the sake of changing? No. I would try to find something that would confirm ... to support. But if I'd change it just for the sake of changing, no." "I may change or may not and how I change it depending on the guy, on the speech sample."* ***OT4:*** *"Yes." "I'd weigh each criterion and consider the overall score from that. I'd average the score in each criterion."* ***OT5:*** *"No." "I mean there"s room for improvement. He could improve. You could have other courses for him to attend, to give him a level four. I mean to actually raise his ability. Not by just giving him a score. I mean he should get what he"s able to do. Right? His ability is only a „three" so he gets a „three". Even if he gets „three" in*

	just one criterion so improve that criterion." "No, I wouldn"t change the scores."

Even though all of them realized the ICAO requirement of giving the lowest score in all six criteria to be the overall score for the candidates, three operational/trained raters (OT2, OT3 and OT4) still admitted that they would change their scores. OT2 said that he would give a candidate his mean score e.g. 4.8 and he even said that he did not like the way ICAO requires. OT3 said that it would depend on his overall impression *"If he was overall „four", „four", „four", „four" with one „three", I'll be more cline to bump that to „four", just because I know that he should be „four" and I know that „four" is minimum requirement"* said OT3. OT4 stated that he would *"weight each criterion and consider the overall score from that"*. OT1 was very hesitant. He said that he did not know and he had to listen to the speech sample for a few more times until he was *"absolutely sure"* in order to *"try to give him four on that"*. OT5 was the only rater in this category who said that he would not change his scores. His perspective was that *"he should get what he"s able to do"* and *"even if he gets „three" in justone criterion so improve that criterion"*.

Table 4.228 shows if the operational/untrained raters (OU) would consider changing the score after knowing that the overall score was based on the lowest score among all six criteria.

Table 4.228: The operational/untrained raters" (OU) consideration to change the score after knowing that the overall score was based on the lowest score among all six criteria

Sub-themes	Rater OU1		Rater OU2		Rater OU3		Rater OU4		Rater OU5		Meaning units
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
53. Consideration of score change after knowing	x		x		x		x			x	***OU1:*** *"I would change from „four" to „five" because his overall was good." "I'll let the overall score overrule the criteria." "His pronunciation may be very bad but he could fly to America without any problem." "He*

that the overall score is based on the lowest score among all six criteria	*understood but his pronunciation might be as bad as „thræ" or „four'" "I believe the overall score is the main consideration because ICAO judges from the overall score."* ***OU2:*** *"I look at the overall. I'll change the score to comply with the overall score that I think he deserves."* ***OU3:*** *"No, I wouldn"t." (After being explained) "Oh! Okay I got it." "It's a case by case basis." (After pondering) "Yes, in this case I consider his overall performance over each individual criterion." "It depends on the individual candidate"s performance." "It also depends on the criteria that I think important. For example, this guy gets „three" in structure and vocabulary which I think they"re not serious."* ***OU4:*** *"I'll stick to the rules." "Well ... If he"s close to „five', I'll give him „five" because I feel that he deserves „five' more than „four'" "I put the overall over each individual criterion." "Personally, I look at the objectives if he can reach that is primary. How he can reach that is secondary. For example, you can get to Chiangmai by any means. You may drive. You may fly. You may even walk. We may consider in details but the objective is to get there. This is the reason why I consider the overall first."* ***OU5:*** *(Long thinking) "If it"s required ..." (Very hesitant) "I gave this guy „thræ" for overall because I thought he deserved that." "I would change this guy"s score to „thræ" according to the rules." "Because it"s the set up*

	standard. I awarded the scores according to the guideline so I"d stick to it."

Four out of five operational/untrained raters (OU1, OU2, OU3 and OU4) accepted that they put the candidates" overall scores as the prime concerns in their decisions. OU1 said that he would *"let the overall score overrule the criteria"* while OU2 would *"change the score to comply with the overall score* that he thought the candidates deserved and OU3 would consider the candidate"s *"overall performance over each individual criterion"*. OU4 seemed to stick to the rules in the beginning before changing his mind later to be *"the overall over each individual criterion"*. OU5 was the sole rater in this batch who said that he would comply with the ICAO standards *"because it"s theset up standard"*.

Table 4.229 shows if the linguistic/trained raters (LT) considered themselves as being harsh or lenient or neither.

Table 4.229: The linguistic/trained raters" (LT) self-consideration as being harsh or lenient or neither

Sub-themes	Rater															Meaning units
	LT1			LT2			LT3			LT4			LT5			
	H	L	N	H	L	N	H	L	N	H	L	N	H	L	N	
52. Self-consideration as being harsh (H), lenient (L) or neither (N)	x					x			x			x		x		***LT1:*** *"I"m a kind of person who „sticks to the rules"." "I don"tmind giving a low score but I do mind giving a high score."* ***LT2:*** *(Laughter) "Harsh? Not harsh and not lenient too." "If he can make it, it"s okay."* ***LT3:*** *"Harsh because, as I said, I have to be neutral." "Well, not harsh. Okay let me change from „harsh" to „as it is"." "Not harsh, not lenient but „as it is", as the*

	evidence shows, based on fact." ***LT4:*** *"Neither because I rate according to what I hear."* ***LT5:*** *"I think I'm lenient because of my concern, as I said."*

LT1 did not clearly answer if she considered herself as „lenient" or „harsh" but *"a kind of person who „sticks to the rules'"*. However, the sentence that *"I don"t mind giving a low score but I do mind giving a high score"* might imply that she was a harsh rater. The other three linguistic/trained raters (LT2, LT3 and LT4) said that they were neither harsh nor lenient, just *„as it is"* (LT3) or *"according to what I hear"* (LT4). LT5 was the only one who admitted that she was lenient.

Table 4.230shows if the linguistic/untrained raters (LU) considered themselves as being harsh or lenient or neither.

Table 4.230: The linguistic/untrained raters" (LU) self-consideration as being harsh or lenient or neither

Sub-themes	Rater															Meaning units
	LU1			LU2			LU3			LU4			LU5			
	H	L	N	H	L	N	H	L	N	H	L	N	H	L	N	
52. Self-consideration as being harsh (H), lenient (L) or neither (N)		x			x			x			x				x	***LU1:*** *"I'm lenient because, as I said, as long as you can communicate it"s fine for me"* ***LU2:*** *"Lenient because if I'm harsh, I'd be stricter than this." "Otherwise I"d focus more on errors they made."* ***LU3:*** *"Lenient because if it"s in between, I'll round it up."*

	LU4: *"Lenient because raters should be able to give the advice to the candidates how they could improve their language proficiency because pilots could do better than this." "I don"t want to see anybody fail." "But if the institution wants to fail some pilots to let them get the language support or the language training, it"s another story."* ***LU5:*** *"It depends on the assessment. It"s clear in case of the objective assessment. In case of subjective assessment, it depends on the reasons to support that." "I think I"mreasonable."*

Four out of five linguistic/untrained raters (LU1, LU2, LU3 and LU4) openly admitted that they were lenient. They were so because for LU1 *"as long as you can communicate it"sfine"*, *"if I"m harsh, I"d be stricter than this"* for LU2, *"if it"sin between, I"ll round it up"* for LU3, and *"I don"twant to see anybody fail"* for LU4. An interesting comment was also made by LU4 as *"but if the institution wants to fail some pilots to let them get the language support or the language training, it"sanother story".* This is another non-linguistic sub-theme that a rater has in his mind which is worth-noting. LU5 was the only one who said he was *"reasonable".* He said, *"in case of subjective assessment, it depends on the reasons to support that"*.

Table 4.231 shows if the operational/trained raters (OT) considered themselves as being harsh or lenient or neither.

Table 4.231: The operational/trained raters" (OT) self-consideration as being harsh or lenient or neither

Sub-themes	Rater															Meaning units
	OT1			OT2			OT3			OT4			OT5			
	H	L	N	H	L	N	H	L	N	H	L	N	H	L	N	
52. Self-consideration as being harsh (H), lenient (L) or neither (N)		x			x				x			x			x	***OT1:*** *"I believe I'd be more lenient than harsh." "Because I feel that I would be giving the benefit if in doubt."* ***OT2:*** *"I think I'm lenient because I know the problems, I heard the problems." "That"s probably the only bias thing I'd have as a rater." "I know their problems as Thai speakers, as being Thais who try to speak English."* ***OT3:*** *"I'm not harsh. I don"t think I'm lenient because ... (laughter)." (Long pause, thinking) "Because for every English test I have the advantage of (?) some other tests but I don"t have the advantage ... (laughter) and the ... yeah ...I"d just like to be fair, you know, not taking advantages from anybody but nobody should take advantages. Just to be fair."* ***OT4:*** *"Not lenient, not harsh because I follow the ICAO guidelines and I do it for the sake of aviation industry. If I tend to either side, what I'm doing will be useless. There should be a standard of this."* ***OT5:*** *"No, just a standard rater, I guess." "Well, if you stick to the standard then you couldn"t say if I was too harsh or I was too lenient, right? ICAO specifies the*

	scales already. So if you stick to the scales then whatever he gets he gets. It doesn"t depend on whether I'm very harsh whether I'm easy. It depends on the rating scales."

Two operational/trained raters (OT1 and OT2) thought that they were lenient because *"I would be giving the benefit if in doubt"* (OT1) and *"I know their problems as Thai speakers, as being Thais who try to speak English"* (OT2). The other three OT3, OT4 and OT5) accepted to be neither harsh nor lenient. OT3 was *"just like to be fair"*. OT4 said that he was to *"follow the ICAO guidelines"* and *"do it for the sake of aviation industry"* which was similar to OT5 who was *"just a standard rater"*.

Table 4.232 shows if the operational/untrained raters (OU) considered themselves as being harsh or lenient or neither.

Table 4.232: The operational/untrained raters" (OU) self-consideration as being harsh or lenient or neither

Sub-themes	Rater															Meaning units
	OU1			OU2			OU3			OU4			OU5			
	H	L	N	H	L	N	H	L	N	H	L	N	H	L	N	
52. Self-consideration as being harsh (H), lenient (L) or neither (N)		x				x	x				x				x	***OU1:*** *"Lenient because I want to see everybody flies." "I'm concerned about safety but it takes time to learn a language and it will ruin everything and it"s impossible to do so." "I'd better let them pass because they"ll come back again in the next three years."* ***OU2:*** *"I'm straightforward because everything has its standard. That"s it if we follow that standard. There"ll be no allegation if we follow the standard which is what I'm worried about."* ***OU3:*** *"Harsh." "You can see from the scores I gave."*

	"I think I have high standard." "I have a definition of „six" among Thai Airways pilots. I compare them with this." "I think ICAO levels are not just about language proficiency measurement. It measures common sense, measures logic. They just use language as a base. There are many native speakers or very high language proficient people who do not have common sense. For example, I ask a guy „which area do you think is the most important?" and he talks about the area in a plane. I mentioned the term „area" as a „topic" but he thinks of an area in a plane instead. This is also a common mistake for native speakers. It's about logic and common sense which even native speakers may not have. People from different parts may have different kind of comprehension and interpretation." ***OU4:*** *"Lenient because I don't wanna see anybody fail." "But it doesn't mean I won't fail anyone. It depends on his performance on that day. I don'twanna be extreme on either side. I always think of myself as a test-taker. What I'd say if I have this kind of test."* ***OU5:*** *"I'm straightforward. Not harsh, not lenient." "Because I don'tknow the test-takers. I just follow the rules."*

Two operational/untrained raters (OU1 and OU4) said that they were lenient because *"I want to see everybody flies"* and *"I'd better let them pass because they'll come back again in the next three years"* (OU1) and *"I don't wanna see anybody fail"* (OU4). The other two (OU2 and OU5) thought they were *"straightforward"* because *"everything has its standard"* (OU2) and *"I don'tknow the test-takers. I just follow the rules"* (OU5). OU3 was the sole rater in this group who thought of himself as „harsh" because *"I think I*

have high standard". In spite of that, he stated some of his non-language-related perspectives as *"I have a definition of „six" among Thai Airways pilots. I compare them with this." "I think ICAO levels are not just about language proficiency measurement. It measures common sense, measures logic. They just use language as a base".*

To summarize the findings of the investigated factors affecting the decision-making of the raters, table 4.233 is illustrated and discussed as follows:

Table 4.233: Summary table of the raters‟ responses to the investigated factors

Factors	LT	LU	OT	OU	Conclusions
1)Educational & rating background	0	0	0	0	These factors did not obviously showed their effects in this study. However, their effects might have been seen through some other factors such as the raters‟ rating strategies, their scoring technique, etc.
2) Mental conditions	0	0	0	0	There was no obvious evidence of this effect on their ratings.
3) Physical conditions	0	0	0	0	Most of the ratings were not affected by this factor.
4) Physical settings	N	N	N	N	This factor did not show any effect on their raters‟ decision-making.
5) Rating strategies	Y	Y	Y	Y	The different groups of raters demonstrated the different strategies of their ratings. Those who were trained employed different strategies from those who were not.
6) Test tasks & speech samples	N	0	0	0	Most of the raters thought that the test tasks were neither too easy nor too difficult. However, some linguistic/untrained raters said that they had no idea if the test tasks in the radiotelephony part were easy or difficult because they did not have knowledge or experience in that while some operational raters

					thought the test tasks were easy for them.
7)Interviewers /interlocutors	N	N	N	N	All raters were satisfied with the interviewers‟ performance.
8) Candidates	Y	Y	Y	Y	The raters in all four groups admitted that they compared a candidate with one or more other candidates.
9) Rating scales & descriptors	Y	Y	Y	Y	Every rater in all four groups showed some certain degrees of difficulties in explaining how they interpreted the scales and descriptors. Most of them did it qualitatively. Still, a few did it quantitatively by actually counting the errors made by the candidates.
10) Cut-off scores	N	Y	N	Y	Some untrained raters, who were unaware of the ICAO required cut-off score of 4 and thought that the candidates‟ ability fell in between two levels such as 3 and 4, gave the score of 3.5 or 3+. This links to the factor of „scoring‟ when they realized the ICAO requirement of a full digit score and had to decide if they would change their awarded score to be full 3 or full 4. this had effect on those raters.
11) Personal relationships	Y	Y	Y	Y	At least one rater in each group admitted that s/he might consider changing her/his score because of her/his personal relationships with the candidates.
12) Scoring	Y	Y	Y	Y	Some of the raters admitted that they would change the scores they already gave to the candidates to conform with the overall scores they thought the

					candidates deserved.
13) Raters‟ harshness/ leniency	0	0	0	0	Most of the raters considered themselves as „not harsh‟ „not lenient‟. This should not imply that this factor does not affect their decision-making because it was just the raters‟ self-consideration.

Y = Yes, it has an effect on the raters‟ decision-making
N = No, it has no effect on the raters‟ decision-making
0 = Not obvious, it is unable to make a conclusion

Based on Table 4.233 above, each factor is discussed as follows:

1. Educational & rating background

Regarding their educational background, as can be seen from the findings (see Table 4.15. 4.16, 4.17, 4.18, 4.19, 4.20, 4.21, and 4.22), all linguistic raters (both trained and untrained) have their academic degrees in English or linguistics or English related i.e. English teaching while none of the operational raters (both trained and untrained) have their academic degrees in English or linguistics or English related. As a result, both groups showed different degrees of familiarity with linguistic terms. In terms of aviation operations and aeronautical communication, they also showed different degrees of familiarity with aviation operations and aeronautical communication, which are parts of the RELTA test tasks. This should have influenced the raters in their ratings because they had different perspectives on English and aviation operations due to their differing degrees of familiarity with both fields. Besides, some linguistic raters, particularly the untrained, stated that they had some degrees of difficulties assessing the terms and vocabularies made by the test-takers in the context of aviation since they did not have sufficient knowledge in that field to judge if the candidates used the correct vocabularies. On the other hand, operational raters (both trained and untrained) accepted that they did not fully understand some linguistic terms e.g. discourse markers, distracting fillers, etc. Raters are required to comprehend these terms since they are mentioned in the ICAO rating scale and descriptors. According to Emery

(2007), these raters may have different perspectives on English due to their differing degrees of familiarity with language and linguistics.

Concerning their rating background, only two linguistic/untrained raters who are Ph.D. students in language assessment and evaluation have fundamental knowledge in the assessment principles; the others do not have direct relationship of proficiency rating with their educational background. Nonetheless, those two linguistic/untrained raters never conducted any kind of proficiency rating before. The trained groups (both linguistic and operational) received their rater training after their graduation. They gained their experience in language assessment and using language descriptors after their rater training courses. Therefore, they may differ in the degrees of experience in language assessment and using language descriptors that may affect their ratings (Emery, 2007).

The trained groups (both linguistic and operational) received their rater training after their graduation. They gained their experience in language assessment and using language descriptors after their rater training courses (see Table 4.23 and 4.25). The linguistic/trained group had the most experience in rating because it was a part of their job with Thai Airways while the operational/trained had less experience in rating because their main job was being interviewers, not raters. The other two untrained groups – both linguistic and operational – had even less experience and knowledge in rating. They also differed in the degrees of using language descriptors (Emery, 2007).

These factors did not obviously show their effects in this study. However, their effects might have been seen through some other factors such as the raters‟ rating strategies (see 5. Rating strategies), their scoring technique (see 12. Scoring).

2. Mental conditions

With regard to busyness, only three raters stated that they were not busy while the others 17 raters said that either they were busy with their family business or their routine jobs/study. This issue may have affected their mental status that consequently influences their ratings. However, only limited evidence (see Table 4.27, 4.28, 4.29, and 4.30) is

available on its effect on rating at the moment. This is an area in need of further comprehensive research and investigation.

According to the Duty Regulations for Crew Members (Thai Airways, 2009) that normally requires a minimum of 24 hour rest period for crew members after their flight duties, any crew who gets rest period less than 24 hours is considered „not having enough rest" and it may affect his fatigue. The same rule may apply to those operational raters who return from their last flight less than 24 hours and have to perform the duty as raters that it may have affected their ratings. Almost all operational raters – trained and untrained – returned from their last flight at least 24 hours.

Only one operational trained rater returned from his last flight in the morning of the rating day (see Table 4.31, 4.32). It means he had less than 24 hour rest period before conducting his duty as a rater. He did not accept that his fatigue affected his ratings, though it might have. This factor has not been studied concerning its effect on rating since this might be the first time that pilots were used as operational raters. Therefore, this area also needs further empirical investigation.

With reference to boredom, exhaustion and/or tiredness on rating, their effects may not be obvious. However, either or all of these may have affected raters" mental condition concerning their concentration in rating (O"Sullivan, 2000 cited in Shaw & Weir, 2007). Raters who get bored during their ratings may tend to lose their concentration, which consequently render them not to fully focus on the ratings.

Almost all raters accepted that they felt bored, exhausted and/or tired during rating (see Table 4.33, 4.34, 4.35, and 4.36). However, none of them felt that it affected their ratings. Still, some might show their loss of concentration, which consequently might render them not to fully focus on the ratings.

On the subject of annoyance, Some kinds of it happened to some raters on the ways to their ratings. Some raters said that it was annoying to a certain extent e.g. a car breakdown or

trying to find of a parking space (see Table 4.37, 4.38, 4.39, and 4.40). However, they did not directly accept that it affected their ratings. These kinds of incidents could make raters feel „irritated" and, as a result, affect their ratings. This factor could not be easily measured since it is internal. But it is irrefutable that it could, up to a certain extent, have influence on raters" decision-making.

From the findings, only limited evidence was available on the effect of rater"s mental conditions on rating in this study because this research was not an experimental one. In fact, 17 raters said that either they were busy with their family business or their routine jobs/study. The findings did not obviously show that it affected the raters" mental status that consequently influenced their ratings. This is also an area in need of further comprehensive research and investigation.

3. Physical conditions

O"Sullivan (2000 cited in Shaw & Weir, 2007) describes short-term ailments as one of the facets affecting rater performance. Though its effect may be subtle, it is worth considering some causes of those short-term ailments. For example, a rater who is allergic to dust from an air-conditioner should not be assigned to do his/her rating in an air-conditioning room. A back pain, which is caused by an unpleasant chair, could be avoided by arranging a rater to sit on a nice and comfortable chair. Raters who have some other kinds of short-term ailments e.g. toothache, cold, etc. should be aware of the possible effects they might have on their ratings.

In respect of short-term ailments, the findings (see Table 4.41, 4.42, 4.43, and 4.44) indicated that almost none of the raters had any kind of short-term ailments during rating. Only one rater said that she was allergic to dust from the air-conditioner while another rater complained of back pain, which was caused by the improper chair. However, both of them did not say that it had any direct effect on their ratings.

Relating to the lack of sleep, its consequences are far more dramatic than being tired in the morning. It can cause drowsiness or even headache. Sleep deprivation can result in

impairment in cognitive function, such as attention, concentration and memory. Lack of sleep can cause mood swings including feeling low or being irritable (Ledoux, 2008). Not getting enough sleep can affect the ability to stay awake during the day or make raters feel fatigued. Life style can also have a huge impact on sleep and sleep quality. For example, irregular bedtimes and wake times might give rise to sleep problems that contribute to sleep deprivation. Operational raters who are pilots flying to different time zones could experience this irregular bedtimes and wake times which may affect their duties as raters. Hence, this is worth being considered when assigning pilots to perform duties as operational raters.

All raters said that they had good sleep the night before rating (see Table 4.45, 4.46, 4.47, and 4.48), except the one (OT4) who just returned in the morning on the rating day (see Table 4.51). However, he said that he thought he had enough rest a few hours at home before coming for rating. This could imply that it did not affect their ratings in this study.

4. Physical settings

Shaw & Weir (2007) mentioned that there has not been any empirical research study concerning the effect of factors associated with the environment or physical setting of the rating process on rater performance. This physical setting could be familiar or unfamiliar to raters e.g. if they are assigned to do their ratings at their office, at home, or some other preferred places. Shaw & Weir stated, "Familiarity with one"s work conditions may result in a more settled and therefore less erratic performance." The provision of air conditioning or the presence of noise could also affect raters. OT4"s remark about „at home rating" that gives no time frame for rating could also give raters more time to „get in depth" compared with rating „on site". This may have effect on the scores they awarded to the candidates.

Regarding the physical settings, this research finding shows that none of the raters had any problem with the room temperature (see Table 4.53, 4.54, 4.55, and 4.56), lighting (see Table 4.57, 4.58, 4.59, and 4.60) and noise (see Table 4.61, 4.62, 4.63, and 4.64) during rating. However, some raters preferred some other places to do their ratings if there were options for them (see Table 4.65, 4.66, 4.67, and 4.68). One rater remarked about „at home rating" that gave no time frame for rating. Other raters said they had more time to „get in-

depth" compared with rating „on site". They commented that this might have effects on the scores they awarded to the candidates.

5. Rating strategies

Concerning rating strategies, different rating strategies may have effect on the scores given by each individual rater. Raters who listen to the speech sample just once from the beginning to the end without stopping to listen to any specific part may overlook some mistakes and tend to focus more, or only, on the overall performance of the test-takers rather than on each criterion. This research result concerning this issue shows that three operational untrained raters used this method of „listening once without stopping" in their ratings. This may be because they are „laypersons" or „non-expert raters" who are "people with no academic training or qualifications in language teaching or testing" (ICAO, 2008: 35). These „laypersons" may incline to pay attention to the overall performance of test-takers. On the other hand, most of the linguistic (both trained and untrained) who have qualifications in language teaching or testing, and the operational trained raters who have training qualifications, used „listening/stopping/note-taking" strategy that enable them to get better chance to spot the test-takers" mistakes.

This finding (see Table 4.69, 4.70, 4.71, and 4.72) revealed that the raters used different strategies in their ratings. Some operational untrained raters used the „listening without stopping" strategy by listening to the speech sample just once from the beginning to the end without stopping to listen to any specific part, which might render them to overlook some mistakes made by the candidates. Most of the linguistic (both trained and untrained) who have qualifications in language teaching or testing, and the operational trained raters who have training qualifications utilized the „listening/stopping/note-taking" strategy that might enable them to get better chance to spot the test-takers" mistakes.

As to the „note-taking" strategy, it is another factor that may affect the awarded scores. Raters who do not take any note during listening to the speech sample tend to miss some or all mistakes made by the test-takers. Missing some mistakes made by the test-takers in some criteria may be a factor leading to error in awarding scores in those criteria. This

research result also showed that the raters who did not take notes at all were in the batch of operational untrained raters who were „laypersons" with no academic training or qualifications in language teaching or testing (see Table 4.80). They claimed that they did not „remember" all mistakes made by the candidates but they looked at the overall performance of the test-takers. This clearly affected their ratings since ICAO requires raters to score each criterion and the lowest score in any criterion is the overall score of that test-taker. Judging the overall score before each criterion is against the way ICAO calls for.

Relating to taking a break, taking any kind of break during rating could have both positive and negative effects in rating. On one hand, raters who do not take any break at all may succeed in keeping their concentrations on their ratings. On the other hand, they may be too tired that they may unintentionally lose their focus. Raters who take a break to go to toilet may get a chance to refresh themselves before returning to do their ratings again while those who take a break to answer phone calls may not be able to keep their attention on ratings. All of these can somehow have an effect on their ratings. The findings disclosed that the raters took some kind of break during rating such as going to toilet, answering phone calls, etc. (see Table 4.81, 4.82, 4.83, and 4.84).

In the matter of stopping the tape, stopping to listen for certain parts could affect the way raters award their scores. The reason is the same as above i.e. those who do not stop at all may miss some mistakes or some good items made by test-takers. Hence, they may give higher or lower scores than they should. This study result showed that most of the operational/untrained raters did not stop to listen for certain parts while the others did (see Table 4.85, 4.86, 4.87, and 4.88). The reason is the same as above, i.e. those who do not stop at all may miss some mistakes or some good items made by test-takers. Hence, they might give higher or lower scores than they should have done. The effect of this technique should also be explained during rater training.

As to the raters" concentration on language or content, or both, Elder"s research finding (1992:15) states that "it is quite conceivable that in assessing use of subject specific language the ESL teachers are focusing on the lexis, grammar and the internal cohesion of

the presentation while the subject specialists are more concerned about the way in which subject content is conceptualized." This finding offers the same evidence as Hadden (1990), Barnwell (1989), Ludwig (1982), and Galloway (1977) in that language experts, whether they are teachers or trained language testers, have different perspectives of second language performance from other „linguistically naïve" native speakers. This may also be the case of this „aviation specific" test of pilots. Those operational, especially untrained, raters who have more expertise in flying and, of course, less in language may put more of their concentration on the test content while the linguistic raters who have less or no knowledge at all in aviation may focus mainly on the language used by the candidates. This may have effect on the scores they give in their ratings. However, the operational/untrained raters in this study seem to be so aware of the purpose of this assessment that none of them concentrated more on the content.

From this research finding, most raters seemed to be aware of the purpose of this language proficiency assessment because they said that they concentrated more on language than the content. However, some of them said that they concentrated equally on both while a few of them stated that they focused on the content (see Table 4.89, 4.90, 4.91, and 4.92).

In connection with focusing on accuracy or fluency, or both, accuracy refers to "the ability to produce grammatically correct sentences" (Richards & Schmidt, 2002: 204). While fluency, which is one of the six criteria required by ICAO in this kind of assessment, refers to "the naturalness of speech production, the degree to which comprehension is impeded by any unnatural or unusual hesitancy, distracting starts and stops, distracting fillers (em…huh…er…) or inappropriate silence" (ICAO, 2004: A-12). Considering from the ICAO requirement that the final score for each test-taker is the lowest among the six criteria, all of these six criteria should be weighted equally. This means fluency should not be taken more or less than the others. Raters who focus more or less on fluency may unintentionally affect the scores they award to test-takers. Still, some raters admitted that they focused more, or less, on fluency than accuracy (see Table 4.93, 4.94, 4.95, and 4.96).

Concerning rating each criterion before or after the overall performance, as required by ICAO that the final score for each test-taker is the lowest among the six criteria (ICAO, 2008), raters should rate each criterion before the overall performance. If they rate the overall performance first, it would affect their final decision because if the overall score they give is higher than the lowest score in any criterion they have two alternatives: one is to lower the overall score to match that lowest score, the other is to raise that lowest score to match the overall score. Therefore, the way raters rate each criterion before or after the overall performance affects their final decision-making. Almost all raters said that they did the right thing by rating each criterion first. In spite of that, some operational raters still rated the overall performance before each criterion (see Table 4.97, 4.98, 4.99, and 4.100).

About the raters‟ concentration on errors, the more raters concentrate on errors made by test-takers, the higher the chance they find those errors, and the lower the score the candidates get. Raters who do so also tend to be harsh or severe, which is another factor accounting for their decision (Lumley & McNamara, 1995; Linacre, 1989; Cason & Cason, 1984). The finding showed that just a few raters concentrated on errors made by the candidates (see Table 4.101, 4.102, 4.103, and 4.104).

As to the types of errors, according to the ICAO requirement that the final score for each test-taker is the lowest among the six criteria (ICAO, 2008); therefore, all six criteria should be determined equally. If raters especially focus on errors in any particular criterion, it may have effect on the final scores given to test-takers. In this study, the results showed that most of the linguistic (both trained and untrained) focused on „the first four‟ criteria which were pronunciation, structure, vocabulary, and fluency while most of operational (both trained and untrained) focused on „comprehension‟ (see Table 4.105, 4.106, 4.107, and 4.108).

On the subject of the raters‟ consideration of the relatedness and the quality of the content, they are not the main focus of this assessment. This is a language proficiency assessment, which evaluates “the ability of test-takers to effectively use appropriate language in operational conditions” (ICAO, 2008: 8). Therefore, raters should focus mainly on the

quality of language produced by test-takers. Otherwise, it may unknowingly affect their ratings. The relatedness/relevance and quality of the content are not the main focus of this assessment. However, this relatedness/relevance and quality of the content was regarded as a part of the criterion of „comprehension" by some raters. Therefore, some of them still admitted that they considered the relatedness/relevance and quality of the content in their ratings (see Table 4.109, 4.110, 4.111, and 4.112).

In respect of the raters" consideration of the candidates" distinctive characteristic, this ICAO required assessment is a criterion-reference assessment, which is defined by Davies et al. (1999: 38) as "a test that examines the level of knowledge of, or performance on, a specific domain of target behaviors (i.e. the criterion) which the candidate is required to have mastered". It is not a norm-referenced assessment (Davies et al: 130), of which their performance is judged with reference to some external criterion other than what stated by ICAO. It means that the scores the test-takers get must be judged, based on those six criteria only. Raters should not consider any other irrelevant subject. Otherwise, those unrelated materials that eventually affect their decision may bias raters. This should be eliminated or reduced in the rater training process. However, this study results show that some trained – both linguistic and operational – raters still considered some of the candidates" distinctive characteristics such as experience, accent, and nationality in their ratings. Some raters seemed to have been influenced by the „halo effect" (Mousavi, 1999: 149) that they considered the candidates" experience or confidence in their ratings. However, the study results showed that some raters, even trained (both linguistic and operational) ones still considered some of the candidates" distinctive characteristics such as experience, accent, and nationality in their ratings. Some raters seemed to have been influenced by the „halo effect" because they considered the candidates" experience or confidence in their ratings (see Table 4.117, 4.118, 4.119, and 4.120).

With reference to putting equal weight on all six criteria, the ICAO requirement for the candidates" overall score is that "the final score for each test-taker should not be the average or aggregate of the ratings in each of the six ICAO language proficiency skills but the lowest of these six ratings" (ICAO, 2008: 19). ICAO (ibid.) states the reason for this is

"because the Operational Level 4 descriptors are developed as the safest minimum proficiency skill level determined necessary for aeronautical radiotelephony communications". This means that raters should put equal weight on all six criteria in their ratings. Otherwise, the scores they award may have been affected by these unequally weighted ratings. It was not surprising that some untrained raters (both linguistic and operational) said that they did not put equal weight on all six criteria in their ratings. What is interesting was that even some operational and one linguistic trained raters accepted that they did not put equal weight in their ratings (see Table 4.121, 4.122, 4.123, and 4.124).

6. Test tasks & speech samples

With regard to test tasks, test task difficulty is one of the factors affecting the scores awarded to test-takers. In Generalizability theory (G-theory), task is considered as a factor or „facet" for specifying and estimating the relative effects of different factors on test scores (the other facets are raters and test-takers) (Upshur & Turner, 1999; Bachman et al., 1995; Bachman, 1990; Brennan, 1983; Cronbach et al., 1972). Rater perspectives on test task difficulties may have effect on rater's decision-making since „easy" tasks may cause raters to be harsher than usual. On the other hand, „difficult" tasks may make raters to be more lenient. Most of the raters in this study should not be biased by the test tasks since they thought the test tasks were neither too easy nor too difficult. However, some operational/trained raters thought that the tasks were „easy" for them (see Table 4.125, 4.126, 4.127, and 4.128). Hence, they may be irritated if the candidates could not perform as they thought the candidates should have been able to do. As a result, this might affect the scores awarded.

Considering speech sample duration, speech sample rating is a time- and energy-consuming job. Raters may get tired and/or bored after doing it for a certain period of time. As mentioned earlier, the effect of boredom, exhaustion and/or tiredness on rating may not be obvious. However, either or all of these may affect raters' mental condition concerning their concentration in rating (O'Sullivan, 2000 cited in Shaw & Weir, 2007). Raters who get bored during their ratings may tend to lose their concentration, which consequently render them not to fully focus on the ratings. As aforementioned concerning mental conditions, the effect of boredom, exhaustion and/or tiredness on rating should be considered before assigning raters

to rate a set of speech samples. The amount of speech samples to be rated in a period of time, e.g. one day, should also be carefully allotted to a rater since this may make them unwittingly tired and/or bored and, consequently, affect their decision-making. In this research finding, most raters said that the rating duration was appropriate though a few complained that it was too long (see Table 4.129, 4.130, 4.131, and 4.132). Most raters suggested that the appropriate duration was between 15-40 minutes (see Table 4.133, 4.134, 4.135, and 4.136).

In terms of an appropriate speech sample duration, due to the effect of boredom, exhaustion and/or tiredness on rating, test administrators should consider the speech sample duration before assigning raters to rate a set of speech samples. The amount of speech samples to be rated in a period of time e.g. one day should also be carefully allotted to raters since this may cause them unwittingly tired and/or bored and, consequently, affect their decision-making. The appropriate number of speech samples the raters suggested was between 3-6 samples a day (see Table 4.137, 4.138, 4.139, and 4.140).

7. Interviewers/ interlocutors

On the subject of interviewers/interlocutors, many researchers have studied the roles of interlocutors in speaking assessment (Brown, 2003; Malvern & Richards, 2002; Jennings et al., 1999; McNamara & Lumley, 1997; Lazaraton, 1996; Ross & Berwick, 1992). Most of the foreign-language speaking assessment uses the oral proficiency interview technique which was developed by ACTFL (American Council on the Teaching of Foreign Languages) OPI proficiency scales out of the FSI (Foreign Service Institute) levels of oral proficiency (ACTFL,1999). This kind of interview technique was criticized concerning the „asymmetric nature" of interlocutor/candidate discourse (Taylor, 2000). It is the interlocutor who leads and controls the interaction during the interview. This creates an imbalance in the power relationship between the interlocutor and the test-taker. However, the effect of the interlocutor in this kind of assessment is undeniable. Various studies show how the behavior of the interlocutor can affect candidate performance (Brown, 2005, 2003; O'Sullivan, 2000; Ross & Berwick, 1992). Brown (2004, 2003) and Brown & Hill (1998) found that raters' perception of a candidate's oral proficiency, which affected the scores they awarded, was influenced by the choice of an interviewer. The results of this study (see Table 4.145, 4.146,

4.147, 4.148, 4.149, 4.150, 4.151, 4.152, 4.153, 4.154, 4.15, and 4.156) show that all raters in all four groups were satisfied with the interviewers‟ performance. Though some of them thought that the interviewers somehow simplified their speech to accommodate or to match the candidates‟ proficiency levels, they eventually admitted that the interviewers performed their jobs appropriately.

In conclusion, the interviewers/interlocutors in RELTA speech samples performed their jobs well enough, so that there was no ill effect from them in this study.

8. Candidates

Concerning the candidates, Candidates/Test-takers can be another factor affecting language test scores. In Generalizability theory (G-theory), test-takers are considered as a factor or „facet‟ for specifying and estimating the relative effects of different factors on test scores (the other facets are raters and test-tasks) (Upshur & Turner, 1999; Bachman et al., 1995; Bachman, 1990; Brennan, 1983; Cronbach et al, 1972). This ICAO language proficiency assessment is a criterion-referenced test (CRT). The test score that each candidate gets reflects that candidate‟s ability in relation to the six criteria, which are evidently stated by ICAO. A candidate‟s performance is not compared with other candidates‟. The factor of candidates/test-takers in this aspect does not mean their language attributes that are measured can affect their scores directly, but it means other non-language related traits that they possess such as their experience and confidence can. Moreover, a candidate‟s performance should not be compared with other candidates‟. Therefore, raters must only rate candidates by comparing the candidates‟ ability with those six criteria, not with other candidates or with other irrelevant factors such as candidates‟ confidence, candidates‟ experience, etc. If the raters compare the candidates or consider other irrelevant factors, they may be biased by those factors as mentioned in some previous findings (Schaefer, 2008; Elder, 1997; Goldstein, 1996; Lumley & McNamara, 1995; Wigglesworth, 1993; Chen & Henning, 1985).

The term “test bias” is closely related to the candidates. It is defined as “any aspect of a test which yields differential predictions for groups of persons distinguishable from each other by a factor which should be irrelevant to the test (Mousavi, 1999: 397). Candidates‟

age, their genders, their global/overall attitudes, and their nervousness are all irrelevant to the test. Raters must not consider these factors in their ratings; otherwise, they will be biased. However, Wigglesworth (1993: 305) stated that the language assessment, particularly speaking and writing, is subjective and "it is subject to the idiosyncratic differences which are found across raters". This idiosyncrasy is arduous to eliminate even after receiving rater training as McNamara (1996: 118) said, "rater differences are reduced by training but do persist". In this study, results show that some raters considered some other irrelevant factors such as raters‟ experience and raters‟ confidence (see Table 4.165, 4.166, 4.167, and 4.168). With reference to the raters‟ sympathy for the candidates, some of the raters even sympathized for the candidates‟ nervousness (see Table 4.173, 4.174, 4.175, and 4.176). In the matter of candidate comparison, the study results (see Table 4.177, 4.178, 4.179, 4.180) show that even some trained raters (both linguistic and operational) admitted that they did compare candidates with others. It was even worse in the untrained group (both linguistic and operational) because almost all of them accepted that they did compare candidates with others.

9. Rating scale & descriptors

A rating scale descriptor is "a statement which describes the level of performance required of candidates at each point on a proficiency scale" (Davies et al., 1999:43). In theory, raters refer to a rating scale in order to select a score to represent the candidate‟s ability in the trait of interest (Upshur & Turner, 1999). In reality, each rater has a unique background that may affect his/her judgment (Brown, 1995; Elder, 1993). Interpretation of a rating scale is always an interest of many researchers. Lumley (1995) found differences in the interpretation of the rating scale used by trained ESL raters and medical practitioners. This finding confirmed Brown‟s study about the perception of language-trained raters and experienced guides in 1995 in that the two groups interpreted different criteria in different ways. Brown‟s conclusion of her study is interesting. She remarked that "raters appear to have inbuilt perceptions of what is acceptable to them and these perceptions are formed to some extent by their previous experience" and "it appears that even the explicitness of the descriptors and the standardization that takes place in a training session cannot remove these differences" (Brown, 1995: 13). The possible implication of this remark is that if the

descriptors are inexplicit, raters‟ perceptions are prone to be based on their previous experience. Imprecise rating scales often results in holistic marking by raters (Weigle, 2002 cited in Knoch, 2009). That leads raters to use the overall or global impression of the candidates in their ratings instead of using an analytic rating scale, as it should be (Knoch, 2009). This misuse has a significant effect on this ICAO proficiency assessment since ICAO requires the lowest score in any criteria to be the overall score (ICAO, 2004). The study results demonstrate that all raters faced a degree of difficulties to explain how they interpreted the ICAO descriptors. Even though it is a common practice in language testing that the descriptors are categorized using adjectives like those mentioned in the ICAO descriptors (Knoch, 2009), each of the raters had dissimilar ideas of the descriptors i.e. „never‟, „almost never‟, „rarely‟, „sometimes‟, „frequently‟ and „usually‟. This is one of the most commonly mentioned problems among raters. They thought that the descriptors were often too vague to arrive easily at a score (Knoch, 2009). This inexplicit interpretation of descriptors by each rater may affect his/her ratings to a certain extent. This is also confirmed by the findings of Eckes (2008) that raters differed significantly in their views on the importance of the various criteria and raters were far from dividing their attention evenly among the set of criteria. Moreover, rater background variables were shown to partially account for the scoring profile differences.

After having difficulties explaining their interpretation of the descriptors, most of the raters seemed to agree that they interpreted the descriptors „qualitatively‟ rather than „quantitatively‟ (see Table 4.185, 4.186, 4.187, and 4.188). Just a few of them said that they did it „quantitatively‟ by counting the frequency of errors made by the candidates. This group of raters seems to try avoiding the subjective scoring, which is always under criticism of its reliability and fairness (Fulcher, 2003; McNamara, 1996). Knoch (2009) also reported that when raters having problems deciding on band levels with the scale, they used various strategies in coping with the problem such as assigning a global (an overall) score, or rating with a halo effect (Lumley, 2002; Vaughan, 1991), or disregarding the descriptors and override them with their own general impression. Moreover, raters may read the descriptors holistically and adjust analytical scores to match their holistic impression (Weigle, 2002 cited in Knoch, 2009).

If raters feel that the descriptors are inexplicitly explained and they are not certain about their interpretations, should they consult the details of those descriptors more frequently? The results reveal that the untrained raters, who were less familiar with the ICAO descriptors than their counterparts were, spent less time consulting the details of the descriptors (see Table 4.189, 4.190, 4.191, 4.192, 4.193, 4.194, 4.195, 4.196, 4.197, 4.198, 4.199, and 4.200).

In spite of "the proven difficulty in defining precisely what a native speaker is" and the ICAO clear statements (ICAO, 2004: 2-9) that its rating scale does not refer to „native" or „native-like" proficiency, together with "there is no presupposition that first-language speakers necessarily conform" (to ICAO rating scale), all raters in this study (both trained and untrained) still provided various reasons when being asked for their opinions to compare between English native speakers and those who are at ICAO level 6 and vice versa instead of quoting what ICAO simply says (see Table 4.201, 4.202, 4.203, 4.204, 4.205, 4.206, 4.207, and 4.208). These raters did not explicitly admit but they might be unaware that they were influenced by the definition of „an educated native speaker" (ILR: 2010a)) and „a well-educated native speaker" (ILR: 2010b) terms, which are used by the United States Foreign Service Institute (FSI) in the Interagency Language Roundtable (ILR) scale to describe the abilities to communicate in a language by ILR level 5. The FSI family of rating scales which is described as „the most influential in the history of testing second language speaking" by Fulcher (2003: 88) since the 1958 FSI band descriptors described its Level 5, its highest level, as „Native or Bilingual" proficiency of which "speaking proficiency is equivalent to that of an educated native speaker" (Fulcher, 2003: 227). It should be emphasized to raters in their rater training that ICAO does not mention anything about „native" or „native-like" in its rating scale descriptors. Otherwise, they may be affected by another irrelevant factor in their ratings.

10. Cut-off score

The perspective of the effect of cut-off score on the raters" decision-making can be viewed through the raters" awareness of the cut-off score and the raters" consideration of the candidates" pass/fail results. McNamara (1996) mentioned some issues in which raters may

differ from one another. One of them is the way they interpret the rating scale they are using. Rating scales usually involve discrete rating categories, which are typically in the range of 1 to 6 such as the ICAO rating scale. The problem arises when raters think that a candidate"s ability falls in between two discrete scales e.g. between Level 3 and 4. McNamara called this kind of situation, into which the rater is forced, an „either/or" judgment (McNamara, 1996: 124). Another issue is how different raters carve the continuum of the rating scale from Level 1 to Level 6. Some raters may carve them with equal intervals while the others may do it unequally. Since ICAO requires that the overall score given to any candidate must be in a full score i.e. not in a decimal or plus/minus e.g. 3.5 or 3+, if raters think that a candidate"s ability falls in between two discrete scales e.g. between Level 3 and Level 4, they are forced to choose between those two levels. If raters realize that the cut-off score is level 4, on one hand they may be prone to the „error of central tendency", which is defined by Mousavi (1999: 143) as "the tendency to avoid all extreme judgments and rate all individuals right down the middle of the scale". On the other hand, raters may be liable to the „generosity error/leniency error" that is "a general tendency for a rater to give every subject the benefit of doubt and when uncertain to give high ratings" (Mousavi, ibid). Both cases result in awarding Level 4 to the candidate. This seemed to be the case of this research finding. All trained raters accepted that they were aware of the ICAO cut-off score as Level 4 while most untrained raters were not (see Table 4.209, 4.210, 4.211, and 4.212). Majority of the operational/trained (four) and the linguistic/trained (three) raters accepted that they considered the candidates" pass/fail results in their ratings while only two raters in each untrained (both linguistic and operational) groups did so (see Table 4.213, 4.214, 4.215, and 4.216).

11. Personal relationships

With respect to the effect of the personal relationships on ratings, since ICAO (2008: 22) requires two kinds of raters, namely linguistic and operational raters, those operational ones are likely to come from the same organization as the candidates. For example, an operational rater who is a pilot working for Thai Airways International may have to assess a candidate who is also a pilot working for Thai Airways International. This is a matter of personal relationships between a rater and a candidate. Both of them may or may not know each other before. If they do, a bias is unavoidable. If they do not, raters may still do not

want to perform the act of „god" because if a candidate does not pass this ICAO required test, his/her license will not meet the full requirements of ICAO regarding the language proficiency. Hence, s/he will be exempted from performing his/her flight duties internationally. This is the „negative appraisal situation" for raters that they may be reluctant to „play god" hence leading to the tendency to be lenient as defensive behavior i.e. avoiding the reactions from candidates who, in this case, are someone with personal relationships by not awarding harsh rating (Bernardin & Buckley, 1981: 209).

In organizations with a system of hierarchical relationships such as in armed forces or civil institutions organized along military lines with regards to logistics, command, and coordination like between captains and flight officers in flight deck operations, language proficiency assessment ratings may be much more complicated than in academic context. The relationship between captains, who are „officially designated leaders" and flight officers, who are their subordinates is clearly described by Kern (2001: 56) that "the cockpit of an aircraft is no place for true democracy" and "the pilot in command has final authority". This authority and superiority of captains are accepted and recognized by flight officers during both line flight operations in flight deck and outside that environment. The problem in pilot language proficiency assessment may arise when flight officers are assigned to be raters and/or interlocutors and they have to rate and/or interview candidates who happen to be their captains. In this case, the raters may be biased or influenced by that hierarchical relationship.

A similar case is reported by Sollenberger (1978: 6) in the early days of the Foreign Service Institute (FSI) Oral Interview Test, which was conducted among military personnel, that "in some cases, the rank and age of the officers were seen to influence the rating". Moreover, "some testers seemed to be unduly influenced by the personalities and cooperativeness of persons being tested".

The results show that most of the operational raters (both trained and untrained) showed some degrees of discomfort when being asked this question (see Table 4.217, 4.218, 4.219, and 4.220). One operational rater even accepted that it had effect on his rating while another suggested that raters should not be people in the same organization as candidates.

While linguistic raters (both trained and untrained) showed less concern about this issue. Just two raters, one each from trained and untrained groups, conveyed their concerns in this matter by considering change of scores because *"it concerns personal relationship"*.

12. Scoring

These two issues of lowest score awareness and score alteration are closely related to the previous issues of raters‟ personal relationships with candidates. Regarding the raters‟ awareness of the overall score that it is based on the lowest score, the research finding (see Table 4.221, 4.222, 4.223, and 4.224). revealed that all trained raters (both linguistic and operational) were aware that the overall score would be considered from the lowest among all six criteria while not all untrained were. Concerning the raters‟ consideration of the score alteration, it is very interesting that the majority of the raters, even the trained ones, admitted that they would change the scores they awarded to the candidates in some criteria to match the overall scores they thought the candidates deserved or to match the cut-off score, i.e. Level 4 (see Table 4.225, 4.226, 4.227, 4.228). Some would award a candidate a higher score if he received higher scores in most criteria. Some said that they had to consider if that lower score was „how low‟ and they had to see *"if that interferes with the overall language or not"*. Some would change the overall to what they had in mind *"because it is overall"*. One rater said that it would depend on his overall impression *"If he was overall „four", „four", „four", „four" with one „thræ", I'll be more cline to bump that to „four"."* because he knew that the candidate *"should be „four"* and he knew that *„four" is minimum requirement"*. All of these ideas do not abide by ICAO requirement that "the final score for each test-taker should not be the average or aggregate of the ratings in each of the six ICAO language proficiency skills but the lowest of these six ratings". This is because "the Operational Level 4 descriptors are developed as the safest minimum proficiency skill level determined necessary for aeronautical radiotelephony communications" and "a lower score than 4 for any one skill area indicates inadequate proficiency. This crucial point must be included and emphasized in the rater training program. Otherwise, it is likely to affect raters‟ decision-making as it happened to the study groups of the untrained raters (both linguistic and operational).

This is not the case of other tests that use the „compensatory composite score" technique which "low scores on a given component/criterion score can be compensated for by high scores on other components/criteria" (Bachman, 2004:318) Those kinds of tests do not require the lowest score in one criterion to be the overall score. Therefore, raters do not have to consider much about the score they would award in each criterion.

13. Rater harshness/ leniency

Rater harshness is one of the factors concerning rater characteristics, which affect the scores awarded to test-takers. In Generalizability theory (G-theory), rater harshness is considered as a factor or „facet" for specifying and estimating the relative effects of different factors on test scores (the other facets are test tasks and test-takers) (Upshur & Turner, 1999; Bachman et al., 1995; Bachman, 1990; Brennan, 1983; Cronbach et al., 1972). McNamara (1996) mentioned that raters may simply differ in their overall harshness/leniency, or they may be consistently lenient on one item while consistently severe on another (rater-item interaction), or they may have a tendency to over- or underrate a candidate or group of candidates (rater-candidate interaction). Even rater training cannot eliminate the extent of rater variability in terms of the overall severity (McNamara: ibid.). However, test administrators and test providers should bear in mind this harshness/leniency factor in the selection and training of raters. This study does not employ the technique such as the Multi-faceted Rasch Measurement (Linacre, 2009; Bond & Fox, 2007; Linacre 1989). This technique has been widely used by many researchers (Knoch, 2009; Eckes, 2005; Kozaki 2004; Kondo-Brown, 2002; Upshur & Turner, 1999; Lynch & McNamara, 1998; Weigle, 1998; Engelhart, 1994; Wigglesworth, 1993; Lunz et al., 1990) to identify the outliers (raters who are too harsh or too lenient) since it was not the objective of this study. The researcher just asked for the raters" self-consideration if they thought they were harsh or lenient to acquire their ideas regarding this topic.

Therefore, it cannot be concluded that this factor has little effect on raters" decision-making. Most of the raters in the three groups thought that they were neither harsh nor lenient. However, the majority of the linguistic/untrained raters thought that they were lenient (see Table 4.229, 4.230, 4.231, 4.232).

There were some other factors which were excluded from the main 13 factors mentioned above. They are summarized for the interest of any further study as follows:

14. Other factors

- Raters" utmost concern in their score awarding

The utmost concern in awarding each candidate"s score for LT1 and LT5 is the same, which is the candidate"s pass or fail. However, the difference between these two linguistic/trained raters is that LT1"s concern is not about his career but it is how she can help to improve his English ability if he fails while LT5 is concerned about the candidates" lives as pass or fail. LT2 had the same concern as LT1 of how she can help to improve the candidate"s English ability if he fails. LT3"s concern is one of the criteria i.e. fluency. She stated how she considered the importance of fluency as taking it prior to accuracy. LT4"s concern is very interesting in that she expressed her skepticism if the score she awarded to the candidate really reflected the candidate"s true ability.

The only linguistic/untrained rater who admitted that he was concerned about the candidates" pass/fail results was LU4. LU2 and LU3 are concerned about the correctness of the scores they give. LU1"s and LU5"s concerns are about the criteria which are comprehension and fluency for LU1 and comprehension and content relatedness for LU5.

Almost all operational/trained raters (OT1, OT2, OT4 and OT5) said that their concerns in awarding the scores is the correctness of the scores if it is the right score for that criterion according to ICAO requirements. OT3 is the sole rater in this category who said that he thought of its consequences that he is *"ruining somebody"s career"* but he finally concluded that it did not have enough effect for him to change his scores.

- Ideal rater characteristics (in the opinions of the raters)

An ideal rater in LT1"s mind *"must be straightforward, honest and knows his job well"*. For LT2 an ideal rater *"must have knowledge, not only in language, but also in assessment", "unbiased",* having *"good listening perception to identify language errors"*

and *"high English competency"*. He/she must also *"be patient and have high concentration"*. LT3"s ideal rater is simply *"straightforward"* and *"neutral"*. LT4"s requirements for an ideal rater are numerous. He/she *"must be so familiar with English"*. That means he/she *"must have very good command of English"* and *"understand the descriptor meaning truly and thoroughly"*. Moreover, an ideal rater *"must be ethical"* and *"strictly adhere to the duties and responsibility as a language assessor"*. LT5 agreed with LT4 in the sense that an ideal rater *"must study the criteria thoroughly and scrupulously"* and uses them *"reliably"*. Even so, LT5 thought that an ideal rater *"must also consider other situation-related factors, not just straight like a ruler. He must be flexible within an acceptable limit."* This was explained as *"for example if he sees that a test-taker has some personal problems, when rating he may consider that the guy may be able to do better if he"s at his 100%. If he"s normal, he"ll have another kind of proficiency but today he has something in his mind that he can"t perform his full competency. When he"s back to normal, he"ll be okay." "The rater may notice from the way the candidate answers the question that he"s not normal."* This explanation seems to reflect the LT5"s perspective as a teacher who always takes care of her students, not just a rater who only assesses candidates.

Unbiased is the common characteristic of an ideal rater among LU1, LU3 and LU5. Other attributes in LU1"view are *"strict"* and *"100% follow the descriptors"* while LU3 thought that he/she *"must understand the criteria thoroughly", "must concentrate on the job during rating"* and *"must be enthusiastic"*. LU5"s ideal rater must also *"be knowledgeable in the field he"s going to assess, must understand the descriptors thoroughly and can differentiate between each level"*. LU2"s perspectives for an ideal rater are just simply *"thoroughly understand the scoring scheme"* and *"should be able to adjust the scales to the actual situation"*. LU4 has a very interesting standpoint of an ideal rater in that on one hand *"in theory, raters must technically follow the scales"*; on the other hand, *"in practice, we must see if we follow the scales, how it would affect their lives"* and *"raters must consider the consequences of the test-takers" scores"*. He clarified this as *"it"s a paradox" "it depends on the institution. What kind of consequences do they want?" I"m personally not a harsh person. I see that you"re still not up to the standard then could you improve yourself? But if the institution says that „get harsh" and it doesn"t matter if they fail, I"d comply with the*

rules. If there"s a kind of training support program for them and the test-takers know that if they fail, they"d be taken care of. They"d be ready to go under that training process. I did the rating today without knowing if there"s such training support program, so I did it generally in the middle. I took balance between two things, the standard and the consequences of the test-takers."

Almost all operational/trained raters (OT1, OT2, OT3 and OT4) seemed to be concerned of the pressure that raters could get in doing this job. OT1"s ideal rater is *"an outsider with aviation background and know the society"* while OT2"s is *"someone who is not affiliated with the organization"* and OT3"s ideal rater is *"not a pilot"*. OT4"s ideal rater is *"brave enough to stand against the outside pressure"*. The other characteristics for them are raters who are *"not having the same nationality as candidates"* and are *"not within the same company as the candidates"* (OT2). For OT3, he/she should be *"somebody knowledgeable in aviation"* and *"fair"*. *"A linguist"* is another requirement for OT3"s and OT4"s ideal rater. An ideal rater should also be *"somebody who got a degree in language and maybe has a PPL as well"* for OT3. An ideal rater should *"have knowledge in language"* and *"unbiased"* for OT4. For OT5 he looked at an ideal rater as a person who has to *"sacrifice by actually spending the time to listen to the overall tape"* and *"has to follow the ICAO standard scales"*.

In OU1"s opinion, an ideal rater should be someone who has *"a lot of experience and unbiased"* while OU2 thought that he/she must be *"straightforward", "strictly adhere to the rules"* and *"consider the consequences of his rating - not in terms of the person but in terms of his job"*. OU3 looked at *"fairness", "good judgment"* and the *"courage to say „no""* as his ideal rater"s attributes. OU4"s perspectives for his ideal rater are *"stick to the rules", "merciful"* and *"understand the nature of Thai people or test-takers"*. One of the OU4"s ideas, which contradict OT1"s, and OT3"s is that *"it"d be even better if he is a pilot"*. OU5 had an opposite perspective from OT2 because he thought that *"he should be a native speaker"*. Nonetheless, OU5 agrees with OT2 and OT3 in that *"he must not be in the same organization as the candidates"*. *"Positive thinking" "unbiased"* and *"unprejudiced"* are additional attributes for an ideal rater in OU4"s and OU5"s opinions.

The term „ideal" may refer to "values that one actively pursues as goals" (Wikipedia, 2010) while „idealization" is defined by Colman (2006: 362) as "a process whereby, the attributes of an instinctual object, especially another person, are represented mentally in a perfected form". The characteristics of a rater in „a perfected form" in the opinions of the raters may somehow reflect the characteristics of raters they possess or, if not, they want to possess. For example, almost all operational/trained raters seemed to be concerned of the pressure that raters could get in doing this job. The ideas such as *"an outsider with aviation background and know the society"*, *"someone who is not affiliated with the organization"*, and someone who is *"brave enough to stand against the outside pressure"* clearly mirror their concerns. The other characteristics such as *„unbiased", „straightforward", „unprejudiced",* and *„fair"* are common attributes mentioned by most raters.

15. Other comments

LT1 thought, *"the third candidate's speech sample is shorter than the others"* while LT2 considered, *"interviewers must be trained too"* because *"they play an important role in this process. Otherwise it will affect raters"*. LT5 thought that *"the interview part which uses plain English should be longer"*. The others in this group did not give any comment.

LU2 thought that the test types and characteristics should have been explained to her in advance and *"there should be a kind of training for raters to clearly understand those terms such as „rarely", „sometimes" or at least give them one day to study the scales and the descriptors before rating"*. LU3 stated her feeling that *"the results may be different between using English language teachers and using aviation experts as raters in this field"*. She also gave a comment that *"people in this field should be raters in this kind of rating, those who have good command of English. Linguists could only rate pronunciation and structure something like that, not the content"*. LU3 considered grammar as *"not so crucial"* because *"if they can communicate correctly, if each party understands each other precisely, that's what we should focus on"*. LU4 admitted that he *"was too tired to write the comments for the third candidate"*. This might imply that he might be too tired to concentrate on his rating of the third candidate. LU1 and LU5 did not make any comment.

OT3 complained that *"the headphone is too tight"* and the screen display while he was listening to the speech samples was *"definitely distracting"*. OT5 suggested that there should be an additional introduction part in the test to let the candidates know what they would face in the test. The other operational/trained raters made no comment.

OU2 expressed his concern about *"the way those who are not raters look at the raters"* in a way that *"they may doubt how good the raters are. How could we explain to them clearly who we are and what kind of standard we have? How could we prove that we are qualified for this job?"* OU4 thought that he should have been given *"more time to study the descriptors"* and *"more direct speech between the interlocutor and the candidates because it"sthe main focus in rating"*. OU5 suggested, *"the descriptors in each level are not clear-cut. They can be interpreted differently by different raters"*. OU1 and OU3 did not give any suggestion.

The subject raters gave comments in various topics concerning this study and others. They are concluded as follows:

- *"Interviewers should be trained too"* (LT2). Many researchers have studied the effects of interviewer/interlocutor in an interview-type test such as the OPI. The studies conducted by Brown (2005, 2003), O"Sullivan (2000b), Lumley & Brown (1996) and Ross (1992) reveal that the behavior of the interviewer has a marked effect on candidate performance. These effects can be controlled, to some certain extent, through training or guided through interlocutor frames (O"Loughlin, 2001; Lazaraton, 1996). McNamara & Lumley (1993) identified three factors regarding the competence an interlocutor should have. First, s/he should have the ability to conduct the test procedure seriously and appropriately. Secondly, s/he should be competent to adopt realistically the role of the simulated person (a patient or a client – in case of McNamara & Lumley"s study or an aviation personnel – in case of ICAO test for aviation) and, thirdly, s/he should be able to establish an appropriate emotional climate, or „rapport", between the participants. Fulcher (2003: 150) mentioned about a good example of interlocutor training by referring

to the advice given to interlocutors in the British Council‟s VOTE: *Oral Testing* video (1983). Some of them are “don‟t correct the test-takers when they make mistakes” and “maintain eye contact with the test-takers”. Test administrators and test providers should consider all of these.

- *“The interview part which uses plain English should be longer”* (LT5). ICAO (2004: 2-3) clearly states “the need for plain language proficiency as a fundamental component of radiotelephony communications”, and its intention to „strengthen‟ the ability to use plain English whenever the ICAO standard phraseologies do not suffice, for example, in emergencies or usual situations. Therefore, the uses of plain English by candidates in speech samples are essential in order for raters to be able to elicit their plain English proficiency. This should be taken into consideration by all test developers and test providers.

- *“There should be a kind of training for raters to clearly understand those terms such as „rarely‟, „sometimes‟ or at least give them one day to study the scales and the descriptors before rating”* (LU2). This comment emphasizes the importance and essence of rater training in view of an untrained rater. Test administrators and test providers should consider these needs and provide them in their rater training course.

- *“The results may be different between using English language teachers and using aviation experts as raters in this field”* and *“people in this field should be raters in this kind of rating, those who have good command of English. Linguists could only rate pronunciation and structure something like that, not the content”* (LU3). ICAO (2008: 22) states that “rating should be carried out by a minimum of two raters” and “ideally, an aviation language test will have two „primary‟ raters – one language expert and one operational expert – and a third rater who can resolve differences between the two primary raters‟ opinion”. ICAO (2008: 46) also defines the term „language rater (or language assessor)‟ as “a rater/assessor whose assessment will evaluate the linguistic features of a test-taker‟s performance in a test”. While „operational rater (or operational assessor)‟ is defined as “a

rater/assessor whose assessment will evaluate not only on the linguistic features of a test-taker‟s performance but also on the appropriateness of a test-taker‟s performance in a test with regard to professional standards and procedures”. It can be seen from the definition that a test should consists of two different kinds of raters to take care of the aspects of both linguistic features and the operational matters. It is worth noting to point out that ICAO does not mention about the qualifications of the „third rater‟ who has the crucial responsibility to „resolve differences between the two primary raters‟ if s/he should be a language or operational rater or, more ideally, should s/he possess both qualifications.

- *“Too tired to write the comments for the third candidate”* (LU4). ICAO (2004: 6-4) requires that raters should not only be able to assess and award scores to candidates but they should also be able to provide „accurate information‟ to candidates who do not pass the test about how their performance fell short of the target performance and in what areas they should focus their efforts to improve their performance. Hence, test administrators should consider managing an appropriate rating duration so that raters would not be too tired to give those „accurate information‟ to candidates. The study results show that linguistic raters – both trained and untrained – seem to give more comments in the rater remark forms compared with the operation raters. This may be because all of them are English teachers who are more familiar with giving comments and feedback to their students.

- *“The headphone is too tight”* and the screen display was *“definitely distracting”* (OU3). These are the problems with the equipment provided for raters to use in their ratings. Test administrators and test providers should consider these „seemingly unimportant‟ things because they may unexpectedly affect the raters‟ ratings.

- OT5‟s suggestion that there should be an additional introduction part in the test to let the candidates know what they would face in the test should be taken into

consideration by test developers and test administrators. This is a matter of test rubrics, which is mentioned by Bachman (1990: 118) as a facet affecting performance on language test.

- OU2 expressed his concern about *"the way those who are not raters look at the raters"* in a way that *"they may doubt how good the raters are"* reflects both the social dimension of the ICAO testing which relates to the face validity of the test, as well as the rater selection process. The rater language proficiency is one of the ICAO"s concerns. It states, "raters should demonstrate language proficiency of at least ICAO Extended Level 5 in the language to be tested." Moreover, "if the test is designed to assess ICAO Level 6 proficiency, raters should demonstrate language proficiency of at least ICAO Expert Level 6" (ICAO, 2008: 36). ICAO additionally explains the reason for these requirements because "test-takers may question the validity and reliability of the test and testing process if they have doubts concerning the credibility and qualifications of the rater". Test developers and test administrators should not overlook the issue of „face validity" too. Though it is just "the degree to which a test appears to measure the knowledge or abilities it claims to measure, as judged by an untrained observer" (Davies et al, 1999: 59). In a case of occupation-specific test like this ICAO required test, "failure to take issues of face validity into account may jeopardize the public credibility of this test" (Davies et al, ibid).

- OU4"s comments that he should have been given *"more time to study the descriptors"* and *"more direct speech between the interlocutor and the candidates because it"s the main focus in rating"* are the same as those of LU2 who emphasized the importance and necessity of rater training in view of an untrained rater. Test administrators and test providers should consider these needs and provide them in their rater training course.

- Slater (1980: 13 cited in McNamara, 1996: 19) pointed out "rating scales with clearly defined levels of achievement can enhance reliability of judge-mediated

ratings." OU5‟s remarks that *"the descriptors in each level are not clear-cut. They can be interpreted differently by different raters"* confirm the needs to train raters to interpret the descriptors in the same way before performing their duties as raters in order to acquire inter-rater‟s reliability. A rater training must focus on the interpretation of the rating scales and descriptors as an essential part of the training.

This chapter describes and discusses the research results. The next chapter summarizes the conclusions and recommendations for future studies.

CHAPTER V
CONCLUSIONS AND RECOMMENDATIONS

Chapter Five presents the research summary, followed by the summary of the findings. It also states the conclusions including implications for the areas of language assessment and evaluation. Subsequently, the recommendations for future research are provided in the last section.

5.1 Research Summary

This study focused on the examination of the two kinds of raters having different background knowledge i.e. linguistic and operational raters with and without rater training experience when they assessed pilots‟ English language speaking performances on RELTA. Furthermore, the other factors affecting their decision-making in awarding the scores to the candidates were explored.

In this study, the RMIT English Language Test for Aviation (RELTA) was used as the data source to obtain the speech samples, which were consequently rated by the raters. RELTA is a standardized test developed by RMIT (Royal Melbourne Institute of Technology) English Worldwide. The test was an early version of RELTA that was conducted with Thai pilots working for Thai Airways International PLC. Three randomly selected of these RELTA speech samples from three different proficiency levels conducted with those Thai pilots were used.

In sum, this study attempts to answer the following four research questions:

1. Does the different background knowledge of raters have any effect on their ratings of Thai pilot speaking ability?
2. Does rater training have any effect on their ratings of Thai pilot speaking ability?
3. Do the different background knowledge of raters and their training have any interactive effects on their ratings of Thai pilot speaking ability?

4. What are other factors affecting the decision making of raters in rating Thai pilot English speaking proficiency?

The participants in this study included 10 operational raters who were Thai pilots from Thai Airways International PLC. The other raters were linguistic raters who were English language teachers. Four of them were from Thai Airways International Flight Crew Language Training Department while the other one was from the Civil Aviation Training Institute. The other five were English language teachers from various institutions. Speech samples from three levels of RELTA (Level 3, Level 4 and Level 5) were used. The other levels (Level 1, Level 2 and Level 6) were excluded because their English proficiency levels are so obviously different that they could easily be distinguished.

RELTA was used as the source of data to obtain the speech samples for the raters to assess. The research instruments consisted of the questionnaire; the rater score sheet and remarks; and the semi-structured interview. The questionnaire for raters was developed primarily from an extensive research of relevant literature (see Chapter 2) and was designed to elicit the raters‟ personal information and ideas, then used with the participants. It was divided into three main parts. Part 1 asked raters‟ general information about their genders, age and educational background. Their experiences in rating were also included in this part. Part 2 investigated the raters‟ strategies used when rating the candidates‟ performance using five-point Likert scale. Part 3 included open-ended questions concerning their opinions about the test and their ratings. The rater score sheet and remarks was provided to each rater in order to specify the scores given to each test-taker in each criterion and the overall score, and to state the reasons why the rater awarded such scores to the test takers or other comments the rater would like to make. Finally, the semi-structured face-to-face interviews were conducted to obtain the in-depth information of the raters‟ strategies towards their ratings of the test takers‟ proficiency and other factors affecting their decision-making.

Because both quantitative and qualitative data were collected to support the findings, this study is the paradigm of mixed methods research (Dornyei, 2007). Qualitative data were

used to supplement quantitative data. To answer the first three research questions and to answer the three hypotheses, the 2x2 ANOVA was employed.

To answer the fourth research question, the content from the interview was grouped into types reported in the literature and was analyzed by the content analysis technique.

5.2 Summary of the Findings

Followings are the hypotheses to answer the first three research questions:

$H^{1}1$: The linguistic raters will rate test takers‟ performance significantly and differently from operational raters ($p \leq .05$).

$H^{1}2$: The raters who are trained in any rater training course will rate significantly and differently from those who are not ($p \leq .05$).

$H^{1}3$: There are significant effects among types of raters, rater training and rating performance ($p \leq .05$);

The results obtained from two-way ANOVA indicated that all hypotheses were rejected. It means that,

- There is no significant difference between linguistic raters and operational raters in rating test takers‟ performance;

- There is no significant difference between trained raters and untrained raters in rating test takers‟ performance;

- There are no significant interaction effects between types of raters and rater training in rating test-takers‟ performance.

In other words, both rater background and rater training did not affect raters‟ decision-making in rating Thai pilots‟ English speaking proficiency, in both main and interaction effects. However, the factor of training seemed to affect the raters‟ decision-making more than the factor of background on the dependent variable.

As for the fourth research question, the content analysis showed that there were 13 factors related to the raters‟ decision-making. These factors were divided into three groups: the factors which had effects on the raters‟ decision-making, the factors which had no effect on the raters‟ decision-making, and the factors which were not obvious, hence, unable to make a conclusion (see Table 4.221). The factors in these three groups are classified based on the results from this study only, which may contradict other previous research studies. They may not be generalized to other contexts as they did not affect or had little effect on raters‟ decision-making since some factors, e.g. the interviewers/interlocutors, the physical setting, were controlled to a certain extent and did not vary so that it was imperceptible to detect if different or varied conditions affected their ratings or not. It can be said that if these factors are well controlled, their affect will be minimized to the least.

5.3 Conclusions

This research study investigates whether there is a significant difference among raters who have different professional backgrounds, i.e. in language and flight operations and those with and without rater training when they evaluate pilots‟ English language performances on RELTA. Quantitative analyses performed show no statistical differences among the four groups in their ratings of candidates‟ oral performances, except in the overall score of speech sample number 3. The differences in raters‟ background (linguistic or operational raters) and their training (trained or untrained raters) had a significant effect in raters‟ rating the “overall” criteria. However, it might not be concluded that, for all criteria, these two factors affect their decision-making in rating Thai pilots' English speaking proficiency. Regarding the qualitative study, content analysis focusing on other factors affecting raters‟ decision-making was conducted. The results were concluded into three groups as follows:

The group of the factors which had effects on the raters‟ decision-making. These factors are rating strategies, candidates/test-takers, rating scale and descriptors, personal relationships between raters and candidates, cut-off score, and scoring.

The group of the factors which had no effect on the raters‟ decision-making. These factors are physical settings, and interviewer/interlocutor.

The group of the factors which were not obvious, hence, unable to make a conclusion. These factors are rater educational and rating background, rater mental conditions, rater‟s physical conditions, test tasks and speech samples, and raters‟ harshness/leniency.

5.4 Implications of the study

The implications from the findings of this study are presented as follows:

1. As for theoretical contribution, the findings provide further insights into the theoretical aspects of the controversial and debatable issue of utilizing different kinds of raters in this high-stakes assessment. The findings of this study substantiate the ICAO requirement of employing two different kinds of raters in this high-stakes test. Since the test is required to be aviation-related, there may be some terms which are specific to the field of aviation that some linguistic raters may not be familiar with. On the other hand, ICAO also requires that "raters should be able to identify deficiencies in performance and guide candidates towards language learning activities that will improve their language proficiency" (ICAO, 2004: 6-4). This finding concurs with ICAO (2004: 6-5) stating that "this is the sort of information that qualified language teachers can provide to candidates". Operational raters may be able to make judgments about language proficiency, particularly in a „pass‟ or „fail‟ sense but they may lack adequate knowledge in linguistics to provide proper information to the candidates on how to improve their language proficiency. This study finding confirms that both kinds of raters are required for this kind of assessment.

2. Concerning pedagogical contribution, rater training, which is the usual means of preparing raters in performance tests that rely on subjective judgments like this, is proved essential for raters to perform their duties properly. Those untrained raters, both linguistic and operational, demonstrated their lack of some vital attributes and knowledge such as the thorough understanding of the ICAO rating scale and descriptors and other requirements e.g. the Level 4 cut-off score.

3. Regarding practical contribution, there are many implications listed below.

3.1 In terms of rater selection and recruitment process in this specific field of aviation, the question of employing just one rater in order to save cost for test administration is open to debate. Another question of using persons who are not in the field of aviation, i.e. linguists and language teachers, to assess language proficiency of pilots and air traffic controllers is also contentious. The results of this study suggest that it may be more costly to use two different kinds of raters (even three raters in case of disagreement between the first two), but it is inevitable because each of them possesses different kinds of expertise, still both types of expertise are essential for this kind of assessment. This study revealed that the linguistic raters who did not have any background in aviation and radiotelephony lacked confidence in rating vocabularies used by the candidates in the context of aviation while the operational raters who did not have background in linguistics were unable to give proper or enough remarks and comments to the candidates. As for the English proficiency improvement required by ICAO, raters should not only award the scores to the candidates but they should also "be able identify deficiencies in performance and guide candidates towards language learning activities" (ICAO, 2004: 6-4). This finding confirms the need of employing both kinds of raters in this ICAO required language assessment.

3.2 ICAO recommends that rating should be carried out by at least two raters, i.e. one linguistic and one operational, and a third rater should be consulted in case of divergent scores (ICAO, 2008). However, ICAO does not clearly state the kind of the third rater. The implication of this study is that the third rater should possess both the aviation operational expertise and the language specialist expertise. That means s/he should be a linguistic and operational rater in one person. As shown in the results of this study, raters with different backgrounds focused on different points.

Most of the operational raters gave priority in their ratings in the criteria of „comprehension" and „interactions". One operational rater stated clearly that he "put more weight on comprehension and interactions" because he gave weight on „comprehensibility/ communicability". Another rater said that he did "not mind much about structure" and "the others such as vocabulary and pronunciation are auxiliary factors". On the contrary, even though they confirmed that they weighted each criterion equally, one of the linguistic raters accepted that she gave importance to the first four criteria. Another acknowledged that she weighted more on the first three criteria i.e. pronunciation, structure, and vocabulary because "they lead to the other three". The third rater should be the one who has knowledge in both linguistics and operations in order to make the final decision in case of disagreement between the linguistic and operational raters.

3.3 ICAO requires that a person"s proficiency rating level is determined by the lowest rating level assigned in any particular category. This research finding revealed that some raters – both trained and untrained – had negative or opposed attitudes towards this, especially those who rated the overall performance of the candidates before each criterion. Even those who rated each criterion first still did not feel „easy" or „comfortable" to change their overall awarded scores to comply with the lowest score in a certain criterion. This specific requirement of ICAO must be reiterated during the rater training session in order to eliminate or attenuate this raters" negative or opposed attitude.

5.5 Recommendations for Future Research

Following are some recommendations for future research:

1. In this investigation, it has been hypothesized that significant differences probably exist among the different groups of raters. This hypothesis is not supported by the findings of this study. Nevertheless, these results should be considered preliminary and should be

augmented with other analyses before any solid implications can be utilized. Even though the quantitative analysis did not show significant differences in terms of the scores awarded to the candidates, the qualitative analysis revealed that the same scores might come from different points of view from different groups of raters. The present research study focused on the use of different kinds of raters in the aviation context. This study is one of the first in this area in Thailand; hence, replications are needed before definite conclusions are made.

2. The present findings are limited by both the speech samples and the raters, i.e. in terms of the small numbers of both groups. Future research should be administered with larger numbers of subjects, both the speech samples and the raters.

3. In addition, the raters included in this study were not necessarily representative of other raters in this industry. Moreover, the analyses undertaken in this study were specific to the operational raters who were pilots working for Thai Airways International and the results may be different if other groups of operational raters were examined, i.e. those from other airlines and those from the field of air traffic controllers. Further investigation should be conducted in the future with other airlines pilots and/or air traffic controllers.

4. This study focused only on rating Thai pilot English speaking proficiency by Thai raters. Further studies should be conducted with some other nationalities, for example using Thai raters to assess other nationalities or using other nationality raters to judge Thai pilots to find out if there are similar factors affecting their ratings.

5. A fundamental issue in this type of assessment is the ability to identify the linguistic and nonlinguistic variables salient to different rater groups when judging pilots‟ English language oral performance. This issue may be the focus of future studies by employing other kinds of investigation like using think-aloud protocols to acquire more in-depth and authentic information from the raters.

6. This study focused only on the information obtained from the rater subjects through the questionnaire and the semi-structured interview. Many factors, e.g. the raters‟

language proficiency, the test difficulty, the raters‟ harshness, etc. were not actually verified. Therefore, some other kinds of analyses may be conducted in the future study such as using the multi-faceted Rasch analysis to investigate the other factors affecting candidates‟ scores, i.e. test difficulty, rater harshness/severity. In terms of raters‟ language proficiency, the rater subjects may be tested to acquire their language proficiency levels or the researchers may use other kinds of evidence such as certificates or scores from some standardized tests e.g. TOEIC, TOEFL, IELTS that the rater subjects possess. Some other kinds of instruments may also be used to obtain data from raters such as the „think-aloud protocol‟ for raters to describe their rating process while they are rating.

7. Some factors such as physical settings, i.e. lighting, noise and temperature were not appropriately controlled in this study. Therefore, it cannot be definitely stated that they do or do not have any effect on the raters‟ decision-making. Future studies should be conducted as experimental research so that these intervening variables can be controlled and varied.

8. In conclusion, the scores obtained from this kind of testing are employed to help make these high-stakes decisions concerning the test takers‟ career and the aviation industry as a whole. It is imperative for all stakeholders to ensure that their tests produce the scores that are supported by research agenda and that they provide quality information in the contexts in which these tests are being implemented. Therefore, test developers and test administrators are responsible for conducting the research needed and undertaking necessary validation research to help make sure that their interpretation and use of test scores are appropriate in this professional context.

References

Aiguo, W. (2007). Teaching aviation English in the Chinese context: Developing ESP theory in a non-English speaking country. *English for Specific Purposes* 26: 121-28.

Alderson, J.C., Clapham, C. & Wall, D. (1995). *Language Test Construction and Evaluation.* Cambridge: Cambridge University Press.

American Council on the Teaching of Foreign Languages. (1999). The *ACTFL Proficiency Guidelines C Speaking*.

Bachman, L.F. (1990). *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Bachman, L.F., Lynch, B.K. & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing* 12: 238-52.

Bachman, L.F. and Palmer, A.S. (1996). *Language Testing in Practice.* Oxford: Oxford University Press.

Barnwell, D. (1989). „Naïve" native speakers and judgments of oral proficiency in Spanish. *Language Testing* 6: 152-63.

Bernardin, H.J. & Buckley, M.R. (1981). Strategies in Rater Training. *The Academy of Management Review* 6 (2): 205-12.

Bond, T.G. & Fox, C.M. (2007). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences.* New Jersey: Lawrence Erlbaum Associates, Inc.

Bordens, K.S. & Abbott, B.B. (2008). *Research Design and Methods: A Process Approach.* New York: McGraw-Hill.

Brennan,R.L. (1983). *Elements of generalizability theory.* Iowa City, IA: The American College Testing Program.

British Council. (1983). *VOTE: Oral Testing.* London: British Council English Languages Services Department and the Design, Production and Publishing Department.

Brown, A. (1995). The effect of rater variables in the development of an occupational-specific language performance test. *Language Testing* 12 (1): 1-15.

Brown, A. (1998). *Interview style and candidate performance in the IELTS Oral Interview*. Paper presented at LTRC, Monterey: March.

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

Brown, A. (2004). Interviewer variability in oral proficiency interviews. Ph.D. thesis. University of Melbourne.

Brown, A. (2005). *Interviewer variability in language proficiency interviews.* Frankfurt: Peter Lang.

Brown, A. & Hill, K. (1998). Interviewer style and candidate performance in the IELTS oral interview. In S. Wood (Ed.) *IELTS Research Reports* 1: 1-19.

Brown, D. (2004). *Language Assessment: Principles and Classroom Practices.* New York: Pearson Education.

Cason, G.J. & Cason, C.L. (1984). A deterministic theory of clinical performance rating. *Evaluation and the Health Professions* 7: 221-47.

Chen, Z. & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing* 2 (2): 155-63.

Cohen, A. (1994). *Assessing Language Ability in the Classroom(7th ed.)* Boston: Heinle & Heinle Publishers.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education.* New York: Routledge.

Colman, A.M. (2006). *A Dictionary of Psychology.* Oxford: Oxford University Press.

Cronbach, L.J., Gleser, G.C., Nanda, H. & Rataratnam, N. (1972). *The dependability of behavioural measurements: theory of generalizability for scores and profiles.* New York: Wiley.

Crystal, D. (1997). *English as an International Language.* Cambridge: Cambridge University Press.

Davies, A. (2001). The logic of testing languages for specific purposes. *Language Testing* 18: 133-47.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. & McNamara, T. (1999). *Studies in Language Testing 7: Dictionary of Language Testing.* Cambridge: Cambridge University Press.

Dornyei, Z. (2007). *Research Methods in Applied Linguistics.* Oxford: Oxford University Press.

Douglas, D. (2000). *Assessing Language for Specific Purposes.* Cambridge: Cambridge University Press.

Douglas, D. (2001). Language for specific purposes assessment criteria: Where do they come from? *Language Testing* 18(2): 171-85.

Downe-Wamboldt, B. (1992). Content analysis: method, applications, and issues. *Health Care for Women International* 13(3): 313-21.

Dudley-Evans, T. & St John, M.J. (1998). *Developments in English for Specific Purposes: A multi-disciplinary approach.* Cambridge: Cambridge University Press.

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.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing,* 25(2): 155-85.

Elder, C. (1992). How do subject specialists construe classroom language proficiency? *Melbourne Papers in Language Testing* 1 (1): 17-33.

Elder, C. (1993). How do subject specialists construe second language proficiency? *Language Testing* 10 (3): 235-54.

Elder, C. (1997). What does test bias have to do with fairness? *Language Testing* 14(3): 261-77.

Elder, C., Brown, A., Grove, E., Hill, K., Iwashita, N., Lumley, T., McNamara, T., & O'Loughlin, K. (ed.). (2001). *Studies in Language Testing 11: Experimenting with uncertainty. Essays in honour of Alan Davies*. Cambridge: Cambridge University Press.

Emery, H. (2006). *Standardization in Language Rating: an Association of Language Raters.* Paper presented at the ICAO Aviation Language Proficiency Seminar, May 1-3, Montreal, Canada.

Engelhard, G.J. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement* 31 (2): 93 – 112.

Ezzy, D. (2002). *Qualitative analysis: Practice and innovation*. London: Routledge

Fulcher, G. (2003). *Testing Second Language Speaking.* Harlow: Pearson Education.

Fulcher, G. and Davidson, F. (2007). *Language Testing and Assessment- An Advanced Resource Book.* Oxon: Routledge.

Galloway, V. (1980). Perceptions of the communicative efforts of American students of Spanish. *Modern Language Journal* 64: 428-33.

George, A.L. (1959). Quantitative and qualitative approaches to content analysis, in Krippendorff, K. & Bock, M.A. (Eds.) *The Content Analysis Reader.* California: Sage: 144-55.

Graneheim, U.H. & Lundman, B. (2004). Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. *Nurse Education Today* 24: 105-12.

Halleck, G.B. (1996). Interrater reliability of the OPI: Using academic trainee raters. *Foreign Language Annals* 29: 223-238.

Hamp-Lyons, L. (1990). Second language writing: Assessment issues, in Kroll, B. (Ed.) *Second Language Writing: Research Insights for the Classroom.* Cambridge: Cambridge University Press: 69-87.

Hamp-Lyons, L. (1991). Reconstructing „Academic Writing Proficiency", in Hamp-Lyons, L. (Ed.) *Assessing Second Language Writing in Academic Context,* Norwood, NJ: Ablex Publishing Corporation: 127-53.

Hadden, B.L. (1991). Teacher and nonteacher perceptions of second language communication. *Language Learning* 41(1): 1-24.

Hinkel, E. (1994). Native and non-native speakers" pragmatic interpretations of English texts. *TESOL Quarterly* 28 (2): 353-76.

Hsieh, H.F. & Shannon, S.E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research,* 15(9): 1277-88.

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

Hughes, A. (2003). *Testing for Language Teachers(2nd ed.)* Cambridge: Cambridge University Press.

Interagency Language Roundtable. (2010a). *ILR scale*. [online] Available from: http://en.wikipedia.org/wiki/ILR_scale. [2010, May 18]

Interagency Language Roundtable. (2010b). *The Interagency Language Roundtable Scale*. [online] Available from: http://www.utm.edu/staff/globeg/ilrhome.shtml [2010, May, 19]

International Civil Aviation Organization. (2001). *Annex 10: Aeronautical Telecommunications. Volume II Communication Procedure (6th ed.)* Montreal: ICAO.

International Civil Aviation Organization. (2004). *Manual on the Implementation of ICAO Language Proficiency Requirements.* Montreal: ICAO.

International Civil Aviation Organization. (2008). *Language Testing Criteria for Global Harmonization.* Montreal: ICAO.

Iwashita, N. (1996). The validity of the paired interview format in oral performance assessment. *Melbourne Papers in Language testing* 5(2): 51-65.

Jennings, M., Fox, J., Graves, B. & Shohamy, E. (1999). The test-takers‟ choice: an investigation of the effect of topic on language-test performance. *Language Testing* 16(4): 426-56.

Kay, M. (2005). *The Development of an ESP Proficiency Test for Civil Airline Pilots: Investigating Construct Validity*. Unpublished Master‟s degree thesis. Melbourne: University of Melbourne.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing* 26(2): 275-304.

Kobayashi, H., & Rinnert, C. (1996). Factors affecting composition evaluation in an EFL context: Cultural rhetorical pattern and readers‟ background. *Language Learning* 46(3): 397-437.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing* 19(1): 3-31.

Kondracki, N.L., Wellman, N.S. & Amundson, D.R. (2002). Content analysis: review of methods and their applications in nutrition education. *Journal of Nutrition Education and Behaviour* 34(4): 224-30.

Kozaki, Y. (2004). Using GENOVA and FACETS to set multiple standards on performance assessment for certification in medical translation from Japanese into English. *Language Testing* 21(1): 1-27.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. California: Sage.

Land, R.E. and Whiteley, C. (1989). Evaluating second language essays in regular composition classes: Towards a pluralistic U.S. rhetoric, in Johnson, D.M. and Roen, D.H. (Eds.) *Richness in Writing: Empowering ESL Students.* New York: Longman: 284-93.

Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: The case of CASE. *Language Testing* 13: 151-72.

Ledoux, S. (2008). *The Effects of Sleep Deprivation on Brain and Behavior.* [online] Available from: http://serendip.brynmawr.edu/exchange/node/1690 [2010, May 11]

Leung, C. and Teasdale, A. (1996). Raters" understanding of rating scales as abstracted concept and as instruments for decision making. *Melbourne Papers in Language Testing* 5(1): 45-71.

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

Linacre, J.M. (2009). *A User"s Guide to FACETS* Rasch-Model Computer Programs. [online] Available from: http://winsteps.com/a/facets.pdf [2010, May 22]

Ludwig, J. (1984). Native speaker judgments of second language learners" efforts at communication: a review. *Modern Language Journal* 66: 274-83.

Lumley, T. (1995). The judgments of language-trained raters and doctors in a test of English for health professionals. *Melbourne Papers in Language Testing* 4(1): 74-98.

Lumley, T. & Brown, A. (1996). Specific purpose language performance tests: Task and interaction. In G. Wigglesworth & /c. /Elder (Eds.), The language testing cycle: From inception to washback. *Australian Review of Applied Linguistics,* Series S (13): 105-36.

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

Lunz. M.E., Wright B.D. & Linacre, J.M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education* 3(4): 331 – 45.

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

Lynch, B.K. & McNamara, T.F. (1998). Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing* 15(2): 158-80.

Madsen, H. (1983). *Techniques in Testing.* Oxford: Oxford University Press.

Malvern, D. & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing* 19(1): 85-104.

McIntyre, P.N. (1993). *The importance and effectiveness of moderation training on the reliability of teacher assessment of ESL writing samples*. Unpublished MA thesis, University of Melbourne.

McNamara, T.F. (1990). *Assessing the language proficiency of health professionals.* Ph.D. thesis, The University of Melbourne.

McNamara, T.F. (1996). *Measuring Second Language Performance.* New York: Addison Wesley Longman Limited.

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

McNamara, T.F. & Lumley, T. (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing* 14(2): 140-156.

Meldman, M.A. (1991). The validation of oral performance tests for second language learners. In M.E. McGroarty and C.J. Faltis (eds.), *Languages in School and Society: Policy and Pedagogy*: 423-438. Berlin: Mouton de Gruyter.

Mousavi, S.A. (1999). *A Dictionary of Language Testing.* (Second edition). Tehran: Rahnama Publications.

O'Loughlin, K. (2001). *The equivalence of direst and semi-direct speaking tests.* Cambridge: Cambridge University Press.

Orr, T. (2002). *English for Specific Purposes.* Virginia: Kirby Lithographic Company.

O'Sullivan, B. (2000a). *Towards a Model of Performance in Oral Language Testing.* Unpublished Ph.D. Dissertation. University of Reading.

O'Sullivan, B. (2000b). Exploring gender and oral proficiency interview performance. *System,* 28: 373-86.

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing* 19(3): 277-95.

Polit, D.F. & Hungler, B.P. (1999). *Nursing research: principles and methods (6th Ed).* Lippincott: Piladelphia.

Porter, D. (1991). Affective factors in the assessment of oral interaction: Gender and status. In S. Anivan (ed.), *Current Developments in Language Testing*: 92-102. Singapore: SEAMEO Regional Language Center.

Reed, D.J., & G.B. Holleck. (1997). Probing above the ceiling in oral interviews: what's up there? In A. Huhta, V. Kohonen, L. Kurki-Suonio and S. Luoma (eds.), *Current Developments and Alternatives in Language Assessment: Proceedings of LTRC 96:* 225-38. Jyvaskyla, Finland: University of Jyvaskyla and University of Tampere.

Reed, D.J., & Cohen. A.D. (2001). Revisiting raters and ratings in oral language Assessment. In *Studies in Language Testing 11: Experimenting with uncertainty. Essays in honour of Alan Davies* by C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara, and K. O'Loughlin. (eds.). Cambridge: Cambridge University Press.

Richards, J.C., & Schmidt, R. (2002). Longman Dictionary of Language Teaching & Applied Linguistics. Third edition. Essex: Pearson Education Limited.

RMIT English Worldwide. (2008). *Aviation English: Evaluation Document.* Melbourne: RMIT.

Robertson, F. (1988). *Airspeak - Radiotelephony Communication for Pilots.* Essex: Pearson English Language Teaching.

Robinson, C. (1993). *Real world research*. Oxford: Blackwell

Ross, S. (1992). Accommodative questions in oral proficiency interviews. *Language Testing* 9: 173-86.

Ross, S., & Berwick, R. (1992). The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition* 14(2): 159-76.

Rubin, H.J. & Rubin, I.S. (1995). *Qualitative Interviewing: The Art of Hearing Data.* California: Sage Publications.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing* 25(4): 465-93.

Shaw, S.D. and Weir, C.J. (2007). *Studies in Language Testing 26: Examining Writing-Research and practice in assessing second language writing.* Cambridge: Cambridge University Press.

Shohamy, E., Gordon, C.M. and Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal,* Spring.

Slater, S.J. (1980). Introduction to performance testing. In Spirer, J.E. (ed.) *Performance testing: issues facing vocational education.* National Center for Research in Vocational Education, Columbus OH, 3-17.

Sunderland, J. (1995). Gender and language testing. *Language Testing Update* 17: 24-35.

Swales, J. M. (1988). *Episodes in ESP.* Hemel Hempstead: Prentice Hall International.

Thai Airways International. (2008). *Duty Regulations for Crew Members.* Bangkok: Thai Airways International.

Underhill, N. (1987). *Testing Spoken Language.* Cambridge: Cambridge University Press.

Upshur, J.A. & Turner, C.E. (1999). Systematic effects in the rating of second-language speaking ability: test method and learner discourse. *Language Testing* 16(1): 82-111.

Valdes, R. (2006). *ICAO Proficiency Requirements in Common English Study Group.* Montreal, Canada, 24-28 April, 2006.

Van Maele, J. (1994). *Native speaker assessment of oral proficiency in Advanced Speakers*. Unpublished manuscript. Brussels: Katholieke Universitieit Brussel.

Vann, R.J., Lorenz, F.O., & Meyer, D.M. (1991). Error gravity: Faculty response to errors in the written discourse of non-native speakers of English, in Hamp-Lyons, L. (Ed.). *Assessing Second Language Writing in Academic Contexts.* Norwood, NJ: Ablex Publishing Corporation: 181-95.

Vaughan, C. (1991). Holistic Assessment: What goes on in the rater‟s mind? In Hamp-Lyons, L. (Ed.) *Assessing Second Language Writing in Academic Contexts*. Norwood, NJ: Ablex Publishing Corporation: 111-25.

Watzlawick, P., Beavin, B.J. & Jackson, D.D. (1967). *Pragmatics of Human Communication. A Study of Interactional Patterns, Pathologies and Paradoxes.* New York: W.W. Norton & Company.

Weigle, S. (1994). Effects of training on raters of ESL compositions. *Language Testing* 11(2): 197-223.

Weigle, S.C. (1998). Using FACETS to model rater training effects. *Language Testing* 15(2): 263-87.

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

Weigle, S.C., Lamison, B. and Peters, K. (2000). *Topic selection on a standardized writing assessment.* Paper presented at Southeast Regional TESOL, Miami, FL, October 2000.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing* 10(3): 305 - 19.

Wikipedia. (2010). *Ideal.* [online] Available from: http://en.wikipedia.org/wiki/Ideal [2010, May 24]

Zhang, Y. & Wildemuth, B. M. (2009). Qualitative analysis of content. In B. Wildemuth (Ed.), *Applications of Social Research Methods to Questions in Information and Library*. Westport: ABC-CLIO.

Appendices

Appendix A

ILR Levels

ILR Level 0 - No functional proficiency.

ILR Level 1 - Elementary proficiency. Elementary proficiency is the first level in the scale. This level is sometimes referred to as S-1 or Level 1. A person at this level is described as follows:

- able to satisfy routine travel needs and minimum courtesy requirements
- can ask and answer questions on very familiar topics; within the scope of very limited language experience
- can understand simple questions and statements, allowing for slowed speech, repetition or paraphrase
- has a speaking vocabulary which is inadequate to express anything but the most elementary needs; makes frequent errors in pronunciation and grammar, but can be understood by a native speaker used to dealing with foreigners attempting to speak the language
- while topics which are "very familiar" and elementary needs vary considerably from individual to individual, any person at the S-1 level should be able to order a simple meal, ask for shelter or lodging, ask and give simple directions, make purchases, and tell time.

ILR Level 2 - Limited working proficiency. This is the second level in the scale. This level is sometimes referred to as S-2 or level 2. A person at this level is described as follows:

- able to satisfy routine social demands and limited work requirements
- can handle with confidence, but not with facility, most social situations including introductions and casual conversations about current events, as well as work, family, and autobiographical information

- can handle limited work requirements, needing help in handling any complications or difficulties; can get the gist of most conversations on non-technical subjects (i.e. topics which require no specialized knowledge), and has a speaking vocabulary sufficient to respond simply with some circumlocutions
- has an accent which, though often quite faulty, is intelligible
- can usually handle elementary constructions quite accurately but does not have thorough or confident control of the grammar.

ILR Level 3 - Professional working proficiency which is the third level in the scale. This level is sometimes referred to as S-3 or Level 3. S-3 is what is usually used to measure how many people in the world know a given language. A person at this level is described as follows:

- able to speak the language with sufficient structural accuracy and vocabulary to participate effectively in most formal and informal conversations on practical, social, and professional topics
- can discuss particular interests and special fields of competence with reasonable ease
- has comprehension which is quite complete for a normal rate of speech
- has a general vocabulary which is broad enough that he or she rarely has to grope for a word
- has an accent which may be obviously foreign; has a good control of grammar; and whose errors virtually never interfere with understanding and rarely disturb the native speaker.

ILR Level 4 - Full professional proficiency. This proficiency is the fourth level in the scale. This level is sometimes referred to as S-4 or level 4. A person at this level is described as follows:

- able to use the language fluently and accurately on all levels normally pertinent to professional needs

- can understand and participate in any conversations within the range of own personal and professional experience with a high degree of fluency and precision of vocabulary
- would rarely be taken for a native speaker, but can respond appropriately even in unfamiliar situations
- makes only quite rare and unpatterned errors of pronunciation and grammar
- can handle informal interpretation from and into the language.

ILR Level 5 - Native or bilingual proficiency. Native or bilingual proficiency is the fifth level in the scale. This level is sometimes referred to as S-5 or level 5. A person at this level is described as follows:

- has a speaking proficiency equivalent to that of an educated native speaker
- has complete fluency in the language, such that speech on all levels is fully accepted by educated native speakers in all of its features, including breadth of vocabulary and idiom, colloquialisms, and pertinent cultural references.

Appendix B

ACTFL Levels

Superior level

Speakers at the Superior level are able to communicate in the language with accuracy and fluency in order to participate fully and effectively in conversations on a variety of topics in formal and informal settings from both concrete and abstract perspectives. They discuss their interests and special fields of competence, explain complex matters in detail, and provide lengthy and coherent narrations, all with ease, fluency, and accuracy. They explain their opinions on a number of topics of importance to them**,** such as social and political issues, and provide structured argument to support their opinions. They are able to construct and develop hypotheses to explore alternative possibilities. When appropriate, they use extended discourse without unnaturally lengthy hesitation to make their point, even when engaged in abstract elaborations. The Superior speakers‟ own language patterns, rather than those of the target language may still influence such discourse, while coherent. Superior speakers command a variety of interactive and discourse strategies, such as turn-taking and separating main ideas from supporting information using syntactic and lexical devices, as well as intonational features such as pitch, stress and tone. They demonstrate virtually no pattern of error in the use of basic structures. However, they may make sporadic errors, particularly in low-frequency structures and in some complex high-frequency structures more common to formal speech and writing. Such errors, if they do occur, do not distract the native interlocutor or interfere with communication.

Advanced High Level

Speakers at the Advanced-High level perform all Advanced-level tasks with linguistic ease, confidence and competence. They are able to consistently explain in detail and narrate fully and accurately in all time frames. In addition, Advanced-High speakers handle the tasks pertaining to the Superior level but cannot sustain performance at that level across a variety of topics. They can provide a structured argument to support their opinions, and they may construct hypotheses, but patterns of error appear. They can discuss some topics abstractly, especially those relating to their particular interests and special fields of expertise, but in

general, they are more comfortable discussing a variety of topics concretely. Advanced-High speakers may demonstrate a well-developed ability to compensate for an imperfect grasp of some forms or for limitations in vocabulary by the confident use of communicative strategies, such as paraphrasing, circumlocution, and illustration. They use precise vocabulary and intonation to express meaning and often show great fluency and ease of speech. However, when called on to perform the complex tasks associated with the Superior level over a variety of topics, their language will at times break down or prove inadequate, or they may avoid the task altogether, for example, by resorting to simplification through the use of description or narration in place of argument or hypothesis.

Advanced-Mid Level

Speakers at the Advanced-Mid level are able to handle with ease and confidence a large number of communicative tasks. They participate actively in most informal and some formal exchanges on a variety of concrete topics relating to work, school, home, and leisure activities, as well as to events of current, public, and personal interest or individual relevance. Advanced-Mid speakers demonstrate the ability to narrate and describe in all major time frames (past, present, and future) by providing a full account, with good control of aspect, as they adapt flexibly to the demands of the conversation. Narration and description tend to be combined and interwoven to relate relevant and supporting facts in connected, paragraph-length discourse. Advanced-Mid speakers can handle successfully and with relative ease the linguistic challenges presented by a complication or unexpected turn of events that occurs within the context of a routine situation or communicative task with which they are otherwise familiar. Communicative strategies such as circumlocution or rephrasing are often employed for this purpose. The speech of Advanced-Mid speakers performing Advanced-level tasks is marked by substantial flow. Their vocabulary is fairly extensive although primarily generic in nature, except in the case of a particular area of specialization or interest. Dominant language discourse structures tend to recede, although discourse may still reflect the oral paragraph structure of their own language rather than that of the target language. Advanced-Mid speakers contribute to conversations on a variety of familiar topics, dealt with concretely, with much accuracy, clarity and precision, and they convey their intended message without misrepresentation or confusion. Native speakers who are unaccustomed to

dealing with non-natives readily understand them. When called on to perform functions or handle topics associated with the Superior level, the quality and/or quantity of their speech will generally decline. Advanced-Mid speakers are often able to state an opinion or cite conditions; however, they lack the ability to consistently provide a structured argument in extended discourse. Advanced-Mid speakers may use a number of delaying strategies, resort to narration, description, explanation or anecdote, or simply attempt to avoid the linguistic demands of Superior-level tasks.

Advanced-Low Level

Speakers at the Advanced-Low level are able to handle a variety of communicative tasks, although somewhat haltingly at times. They participate actively in most informal and a limited number of formal conversations on activities related to school, home, and leisure activities and, to a lesser degree, those related to events of work, current, public, and personal interest or individual relevance. Advanced-Low speakers demonstrate the ability to narrate and describe in all major time frames (past, present and future) in paragraph length discourse, but control of aspect may be lacking at times. They can handle appropriately the linguistic challenges presented by a complication or unexpected turn of events that occurs within the context of a routine situation or communicative task with which they are otherwise familiar, though at times their discourse may be minimal for the level and strained. Communicative strategies such as rephrasing and circumlocution may be employed in such instances. In their narrations and descriptions, they combine and link sentences into connected discourse of paragraph length. When pressed for a fuller account, they tend to grope and rely on minimal discourse. Their utterances are typically not longer than a single paragraph. Structure of the dominant language is still evident in the use of false cognates, literal translations, or the oral paragraph structure of the speaker's own language rather than that of the target language. While the language of Advanced-Low speakers may be marked by substantial, albeit irregular flow, it is typically somewhat strained and tentative, with noticeable self-correction and a certain grammatical roughness. The vocabulary of Advanced-Low speakers is primarily generic in nature. Advanced-Low speakers contribute to the conversation with sufficient accuracy, clarity, and precision to convey their intended message without misrepresentation or confusion, and native speakers who are unaccustomed to dealing with

non-natives can understand it, even though this may be achieved through repetition and restatement. When attempting to perform functions or handle topics associated with the Superior level, the linguistic quality and quantity of their speech will deteriorate significantly.

Intermediate-High Level

Intermediate-High speakers are able to converse with ease and confidence when dealing with most routine tasks and social situations of the Intermediate level. They are able to handle successfully many uncomplicated tasks and social situations requiring an exchange of basic information related to work, school, recreation, particular interests and areas of competence, though hesitation and errors may be evident. Intermediate-High speakers handle the tasks pertaining to the Advanced level, but they are unable to sustain performance at that level over a variety of topics. With some consistency, speakers at the Intermediate High level narrate and describe in major time frames using connected discourse of paragraph length. However, their performance of these Advanced-level tasks will exhibit one or more features of breakdown. These may be the failure to maintain the narration or description semantically or syntactically in the appropriate major time frame, the disintegration of connected discourse, the misuse of cohesive devises, a reduction in breadth and appropriateness of vocabulary, the failure to successfully circumlocute, or a significant amount of hesitation. Intermediate-High speakers can generally be understood by native speakers unaccustomed to dealing with non-natives, although the dominant language is still evident (e.g. use of code-switching, false cognates, literal translations, etc.), and gaps in communication may occur.

Intermediate-Mid Level

Speakers at the Intermediate-Mid level are able to handle successfully a variety of uncomplicated communicative tasks in straightforward social situations. Conversation is generally limited to those predictable and concrete exchanges necessary for survival in the target culture; these include personal information covering self, family, home, daily activities, interests and personal preferences, as well as physical and social needs, such as food, shopping, travel and lodging. Intermediate-Mid speakers tend to function reactively, for example, by responding to direct questions or requests for information. However, they are capable of asking a variety of questions when necessary to obtain simple information to

satisfy basic needs, such as directions, prices and services. When called on to perform functions or handle topics at the Advanced level, they provide some information but have difficulty linking ideas, manipulating time and aspect, and using communicative strategies, such as circumlocution. Intermediate-Mid speakers are able to express personal meaning by creating with the language, in part by combining and recombining known elements and conversational input to make utterances of sentence length and some strings of sentences. Their speech may contain pauses, reformulations and self-corrections as they search for adequate vocabulary and appropriate language forms to express themselves. Because of inaccuracies in their vocabulary and/or pronunciation and/or grammar and/or syntax, misunderstandings can occur, but sympathetic interlocutors who are accustomed to dealing with non-natives generally understand Intermediate-Mid speakers.

Intermediate-Low Level

Speakers at the Intermediate-Low level are able to handle successfully a limited number of uncomplicated communicative tasks by creating with the language in straightforward social situations. Conversation is restricted to some of the concrete exchanges and predictable topics necessary for survival in the target language culture. These topics relate to basic personal information covering, for example, self and family, some daily activities and personal preferences, as well as to some immediate needs, such as ordering food and making simple purchases. At the Intermediate-Low level, speakers are primarily reactive and struggle to answer direct questions or requests for information, but they are also able to ask a few appropriate questions. Intermediate-Low speakers express personal meaning by combining and recombining into short statements what they know and what they hear from their interlocutors. Their utterances are often filled with hesitancy and inaccuracies as they search for appropriate linguistic forms and vocabulary while attempting to give form to the message. Their speech is characterized by frequent pauses, ineffective reformulations and self-corrections. Their pronunciation, vocabulary and syntax are strongly influenced by their first language but, in spite of frequent misunderstandings that require repetition or rephrasing, Intermediate-Low speakers can generally be understood by sympathetic interlocutors, particularly by those accustomed to dealing with non-natives.

Novice-High Level

Speakers at the Novice-High level are able to handle a variety of tasks pertaining to the Intermediate level, but are unable to sustain performance at that level. They are able to manage successfully a number of uncomplicated communicative tasks in straightforward social situations. Conversation is restricted to a few of the predictable topics necessary for survival in the target language culture, such as basic personal information, basic objects and a limited number of activities, preferences and immediate needs. Novice-High speakers respond to simple, direct questions or requests for information; they are able to ask only a very few formulaic questions when asked to do so. Novice-High speakers are able to express personal meaning by relying heavily on learned phrases or recombination of these and what they hear from their interlocutor. Their utterances, which consist mostly of short and sometimes incomplete sentences in the present, may be hesitant or inaccurate. On the other hand, since these utterances are frequently only expansions of learned material and stock phrases, they may sometimes appear surprisingly fluent and accurate. These speakers‟ first language may strongly influence their pronunciation, as well as their vocabulary and syntax when they attempt to personalize their utterances. Frequent misunderstandings may arise but, with repetition or rephrasing, sympathetic interlocutors who are used to non-natives can generally understand Novice-High speakers. When a Novice- High speaker is called on to handle simply a variety of topics and performs functions pertaining to the Intermediate level, s/he can sometimes respond in intelligible sentences, but will not be able to sustain sentence level discourse.

Novice-Mid Level

Speakers at the Novice-Mid level communicate minimally and with difficulty by using a number of isolated words and memorized phrases limited by the particular context in which the language has been learned. When responding to direct questions, they may utter only two or three words at a time or an occasional stock answer. They pause frequently as they search for simple vocabulary or attempt to recycle their own and their interlocutor‟s words. Because of hesitations, lack of vocabulary, even sympathetic interlocutors who are accustomed to dealing with non-natives may understand inaccuracy, or failure to respond appropriately, Novice-Mid speakers with great difficulty. When called on to handle topics by

performing functions associated with the Intermediate level, they frequently resort to repetition, words from their native language, or silence.

Novice-Low Level

Speakers at the Novice-Low level have no real functional ability and, because of their pronunciation, they may be unintelligible. Given adequate time and familiar cues, they may be able to exchange greetings, give their identity, and name a number of familiar objects from their immediate environment. They are unable to perform functions or handle topics pertaining to the Intermediate level, and cannot therefore participate in a true conversational exchange.

Appendix C
Questionnaire

Please indicate your selected items by circling a number and fill in the blank, if applicable.

1. Gender
1) Male
2) Female

2. Age in years
1) 21 – 30
2) 31 – 40
3) 41 – 50
4) 51 – 60

3. Educational level
1) Bachelor"s degree in _____
2) Master"s degree in _____
3) Doctoral degree in _____
4) Others in_____

4. Occupation
1) Linguist/Language teacher
2) Pilot
3) Air traffic controller

5. Years of being in the occupation
1) 1 – 5
2) 6 – 10
3) 11 – 15
4) 16 – 20
5) 21 – 25
6) Over 25

6. Your first language (L1)
1) Native speaker of English (Skip to Item 9)
2) Non-native speaker of English
(Please specify your native language) _____

7. Duration of your English study (If you are a non-native speaker of English)
1) Less than 10 years
2) 11 – 15 years
3) More than 16 years

8. How do you consider your level of English proficiency (If you are a non-native speaker of English)?

1) Native-like/Near native
2) Very good
3) Good
4) Fair

9. Have you been specifically trained in any formal rater training program?

1) Yes (Please specify the name, place and time of the program)

   _____
   _____
   _____

2) No

Please select your answer by putting a ✓in an appropriate space

No.	Questions	1 None	2 Little	3 Some	4 Much	5 Very much
10.	To what extent are you familiar with various English native speakers" accents (e.g. American, British, Australian, etc.)?					
11.	To what extent are you familiar with Asian English accents (e.g. Thai, Chinese, Japanese, etc.)?					
12.	To what extent are you familiar with European English accents (e.g. French, German, Spanish, Scandinavian, etc.)?					
13.	To what extent are you familiar with linguistic terms i.e. those stated in ICAO Doc. 9835 e.g. discourse markers, connectors, stylistic device, etc.?					
14.	To what extent are you familiar with aviation operations and aeronautical communication and the terms used in the aviation context e.g. runway incursion, low pass, etc.?					

No.	Questions	1 Never	2 Rarely	3 Some-times	4 Frequently	5 Always
15.	To what extent are you familiar with language assessment?					
16.	To what extent are you familiar with the use of language descriptors in language assessment?					
17.	How do you consider your familiarity with ICAO language proficiency rating scale?					
18.	How often do you consult the details of each ICAO descriptor in Doc. 9835 before listening to the speech samples?					
19.	How often do you consult the details of each ICAO descriptor in Doc. 9835 during listening to the speech samples?					
20.	How often did you consult the details of each ICAO descriptor in Doc. 9835 after listening to the speech samples?					
21.	How many times did you listen to the given speech samples before giving the final score?					
22.	How often did you take notes while rating?					
23.	How often did you stop the tape for any reason while rating?					
24.	How often did you stop to listen for certain parts from the speech samples?					
25.	How often did you concentrate on errors made by the speaker?					
26.	How often did you consider the relatedness/relevance of the content as a factor in your rating?					

		Yes	No
27.	Have you been busy lately?		
28.	Do you feel bored/exhausted/tired during your rating?		
29.	During rating, do you have any short term ailments such as temporary aches and pains e.g. toothache, headache, back pain, etc. or cold, flu, diarrhea or allergies, etc.?		
30.	During rating, do you have any long term ailments such as speaking, hearing, vision impairment, heart disease, sinus, diabetes?		
31.	Did you have a good sleep/rest last night?		
32.	Do you think you had enough sleep/rest?		
33.	Was the room you did your rating too cold?		
34.	Was the room you did your rating too warm?		
35.	Was the room too dark?		
36.	Was the room too lighted?		
37.	Was the room too noisy?		
38.	Did you listen to the given speech sample from the beginning to the end without stopping at least once before rating?		
39.	Do you think you weighted each criterion equally before giving the final score?		
40.	Do you consider the quality of the content the candidates give as a factor in your rating?		
41.	Do you think the test tasks were easy?		
42.	Do you think the test tasks were difficult?		
43.	Do you think the speech samples were too short?		
44.	Do you think the speech samples were too long?		

		Yes	No
45.	Do you think that rating three speech samples consecutively was too much?		
46.	Do you think any of the interviewers/interlocutors tried to help/accommodate the candidate during the test?		
47.	Do you think the interviewers/ interlocutors performed their jobs appropriately/effectively as they should have?		
48.	Do you think the interviewers/ interlocutors attempted to simplify their speech to facilitate the candidates or to match the candidates‟ level of language?		
49.	Did you consider the candidates‟ age in your rating?		
50.	Did you consider the candidates‟ gender in your rating?		
51.	Did you consider the global/overall attitudes of the candidates?		
52.	Do you think the candidates were nervous during the test?		
53.	Did you sympathize for that in your rating?		
54.	Did you compare the candidate with other candidates in your rating?		
55.	Do you think that every English native speaker must also be ICAO Level 6?		
56.	Do you consider being ICAO Level 6 equivalent to being an English native speaker?		
57.	Do you know that the „cut-off‟ score for this ICAO assessment is level 4?		
58.	Did you consider the consequence of the candidates as being passed or fail in your rating?		
59.	If you have any relationship with the candidates e.g. being friends or relatives, do you consider changing the scores you already gave them?		
60.	Do you consider yourself as being a lenient rater?		
61.	Do you consider yourself as being a harsh rater?		

Other comments

_____

_____

Thank you very much for your cooperation.

Sutas Dejkunjorn

Appendix D
Interview questions

Semi-structured interviews

Rater: _____

1. **What was your major when you studied in the university in both undergraduate and post-graduate levels (if applicable)?**
2. **Was it related to rating? How?**
3. **Have you been busy lately?**
4. **When did you return from your last flight? (For operational raters)**
5. **Do you feel bored/exhausted/tired during your rating? Why?**
6. **During rating, do you have any short term ailments such as temporary aches and pains e.g. toothache, headache, back pain, etc. or cold, flu, diarrhea or allergies, etc.?**
7. **Did you have a good sleep/rest last night?**
8. **Do you think you had enough sleep/rest last night?**
9. **Was there any incident happened to you on the way to your office that annoyed you e.g. a car accident?**
10. **Was the room you did your rating too cold or too warm?**
11. **Was the room too dark or too lighted?**
12. **Was the room too noisy?**
13. **What kind of physical setting do you like when you do your rating?**
14. **How many times did you listen to the given speech before giving the final score?**
15. **How often did you take notes while rating?**
16. **Did you listen to the given speech sample from the beginning to the end without stopping at least once before rating?**
17. **How often did you stop the tape for any reason while rating?**
18. **How often did you stop to listen for certain parts from the speech sample?**
19. **Did you concentrate on the content first or on the language first? / Which do you consider a priority, content or language?**

____ 1) Content first

____ 2) Language first

20. Did you focus on accuracy or fluency or both?

/Which is the most important: accuracy, fluency or both?

____ 1) Accuracy

____ 2) Fluency

____ 3) Both

21. How did you rate each criterion before or after rating the overall performance?

____ 1) Rate each criterion first, then rate the overall performance.

____ 2) Rate the overall performance first, then rate each criterion.

22. How often did you concentrate on errors made by the speaker?

23. What kinds of errors did you listen for?

____ 1) Pronunciation

____ 2) Structure

____ 3) Vocabulary

____ 4) Fluency

____ 5) Comprehension

____ 6) Interactions

____ 7) Others (Please specify _____)

24. How often did you consider the relatedness/relevance of the content as a factor in your rating?

25. How often do you consider the quality of the content the candidates give as a factor in your rating?

26. Is there any distinctive characteristic of the candidate you particularly consider? If so, what are they?

____ 1) Accent

____ 2) Voice

____ 3) Tone

____ 4) Nationality

____ 5) Others (Please specify _____)

27. **Do you think you weighted each criterion equally before giving the final score? If not, why?**
28. **What do you think about the test task, easy or difficult? Why?**
29. **Do you think the speech samples were too short or too long?**
30. **What should be the appropriate duration/length in your opinion?**
31. **Do you think that rating three speech samples consecutively was too much?**
32. **If so, in your opinion, how many speech samples should be rated at one time/in one day?**
33. **Do you think any of the interviewers/interlocutors who tried to help/accommodate the candidate during the test? What made you think so?**
34. **Do you think the interviewers/interlocutors performed their jobs appropriately/effectively as they should have?**
35. **Do you think the interviewers/interlocutors attempted to simplify their speech to facilitate the candidates or to match the candidates" level of language?**
36. **Did you consider the candidate"s age in your rating?**
37. **Did you consider the candidate"s gender in your rating?**
38. **Did you consider the global/overall attitudes of the candidate?**
39. **Do you think the candidates were nervous during the test?**
40. **If so, did you sympathize for that in your rating?**
41. **Did you compare the candidate with other candidates in your rating? Why?**
42. **How do you consider your familiarity with the ICAO language proficiency rating scale?**
43. **When using the ICAO rating scale descriptors, how do you interpret the descriptors? i.e.**

    "almost never interfere with ease of understanding" in pronunciation

    "rarely interfere …"

    "only sometimes interfere …"

    "frequently interfere …"

    "usually interfere …"
44. **Do you consider these descriptors quantitatively or qualitatively? Please explain.**

45. **How often do you consult the details of each ICAO descriptor in Doc. 9835 before listening to the speech samples?**
46. **How often do you consult the details of each ICAO descriptor in Doc. 9835 during listening to the speech samples?**
47. **Did you consult the details of each ICAO descriptor in Doc. 9835 after listening to the speech samples?**
48. **Do you think that every English native speaker must also be at ICAO Level 6? Why?**
49. **Do you consider being at ICAO Level 6 equivalent to being an English native speaker? Why?**
50. **Do you know that the „cut-off" score for this ICAO assessment is level4?**
51. **If so, did you consider the consequence of the candidate as being pass or fail in your rating? Why?**
52. **If you have any relationship with the candidates e.g. being friends or relatives, do you consider changing the scores you already gave them? Why?**
53. **Do you know that ICAO requires the overall score to be based on the lowest score among all six criteria?**
54. **After knowing, would you consider changing your given score? If so, how would you change it?**
55. **Do you consider yourself as being lenient or harsh? Why?**
56. **What is your utmost concern in awarding each candidate"s score?**
57. **What are the characteristics of your ideal rater?**
58. **Do you have any other comments to make?**

Appendix E

Rating score sheet & remarks

Pilot Language Proficiency Rating

Test taker ID	
Test Date	
Rater	

Individual Ratings and Final Rating	
Pronunciation	
Structure	
Vocabulary	
Fluency	
Comprehension	
Interactions	
ICAO Language Proficiency Level (Lowest among individual)	

General Remarks

_____

_____

_____

Rater"s signature _____

( )

Individual Remarks

Pronunciation

_____

_____

____

Structure

_____

_____

____

Vocabulary

_____

_____

____

Fluency

_____

_____

____

Comprehension

_____

_____

____

Interactions

_____

_____

____

Appendix F

Raters‟ remarks on speech sample number 1

Pronunciation

Groups of Raters (Scores)	Remarks

OT1 (4): Pronunciation was good for most part. He had good rhythm and made proper stresses on his vocabulary. (There is) only little interference with understanding.

OT2 (5): Some problems with clusters, final sounds and intonation. Thai accent is present but rarely interferes with pronunciation.

OT3 (4): Pronunciation is influenced by first language but mostly doesn‟t interfere with ease of understanding.

OT4 (4): Problem with “r” and “l” e.g. saying “full range” instead of “full length”. Speak without final sound e.g. “behine” instead of “behind”, “decimon” instead of “decimal”. Leave “l” sound in words such as “holding”, “complicate”. Intonation and pronunciation are influenced by the first language e.g. “koppy” instead of “copy”.

OT5 (4): (He is) influenced heavily by (his) mother tongue but rarely interfere with meaning of the words.

OU1 (5): Some words are unclearly pronounced but can be understood.

OU2 (5): Sometimes obviously influenced by heavy accent but it is acceptable for non-native English speaker.

OU3 (4): Clear and articulate, with good tempo (not to fast) – definitely a strong point.

OU4 (5): Mostly understandable with minor errors in some unfamiliar words or sentences due to his strong Thai accent. Usual practice in pronunciation will improve his ability to demonstrate nearly perfect English speaking.

OU5 (4): Influenced by first language. Occasional interference with understanding.

LT1 (5): He has some Thai interference and still speaks with the problems of final sounds and endings omission. There are some mispronouncing words as well. However, he can be understood most of the time.

LT2 (4): Thai interference mostly on suprasegmental features. Quite numbers of mistakes on word stress and final sounds. However, the overall speech is rather comprehensible as the context of talk is very confined.

LT3 (4): His pronunciation, stress, rhythm, and intonation are influenced by the first language but only sometimes with ease of understanding. Continuing practice on final sounds (v, nce, st, ed), th, r and l sounds, clusters with r sounds, and adding a little more stress and intonation can help him successfully achieve a higher level.

LT4 (4): ID 1‟s pronunciation is somewhat influenced by his L1 especially some final sounds e.g. decimal (decimon), length (lenge), side of taxiway (sign of taxiway), malfunction (more function), words containing „th‟ (/ /, / /) are also difficult for him to pronounce but these do not interfere with ease of understanding as his speech rhythm and intonation are well controlled.

LT5 (5): Influenced by Thai language but easy to understand. (She also jotted down some pronunciation errors that the test-taker made e.g. no final „l" sound such as „decimon", no final „t" sound such as „aircraf", „reques", „jus".)

LU1 (4): It is clear that pronunciation, stress, rhythm, and intonation are influenced by the first language. However, it is understandable and clear when operating.

LU2 (3): His pronunciation is influenced by the first language, so he doesn't show any correct stress and intonation. In addition, he doesn't speak with clear voice.

LU3 (4): Pronunciation generally doesn't interfere with ease of understanding or meaning though it is clearly influenced by L1. The control of pronunciation is better when the test taker feels more relaxed especially in the last section of the test.

LU4 (5): Pronunciation is generally clear, however, it is obvious to guess which country the pilot comes from. Word pronunciation can be improved by practicing e.g. the word "decimal" is heavily influenced by Thai as the pilot's first Language.

LU5 (4): His pronunciation is acceptable. The pronunciation, rhythm, and intonation (are) still influenced by his first language, Thai. However, it does not much interfere (with) the case of understanding.

Structures

Groups of Raters (Scores)	Remarks

OT1 (4): During normal operation the subject did very well. But when he was pushed out of his normality, his structure broke down e.g. FOD and bomb.

OT2 (4): Grammatical errors are more frequent but do not distort meaning.

OT3 (4): Some errors may occur when putting together sentences but they do not interfere with the meaning.

OT4 (4): (Incorrect) word form e.g. "to higher rate", "expect bomb", "no time for inform". (Incorrect) plural forms e.g. "year" instead of "years", "three engine" instead of "three engines". (Wrong) tense e.g. "use to flew", "I lost my job" instead of "I'll lose my job".

OT5 (4): Short and simple sentences are well structured. (He is) not able to make complex sentences. When emergency arises, tenses are not well controlled.

OU1 (5): Understandable sentences with some errors in grammar e.g. "Due to we have warning …"; "I don't sure."; "I will taxi with careful."; "I used to flew."

OU2 (4): Have a god command for basic grammatical structures but for complex situation may demonstrate particularly local errors.

OU3 (3): Somewhat weak on grammar which sometimes gets in the way of describing a situation.

OU4 (4): The sentence structures are mostly based on Thai language structures. However, English native speakers will definitely understand the meanings he wants to get through. The use of sentences is based on translation from Thai to English which leads to usual grammatical errors.

OU5 (4): Mainly well controlled with errors when confronted by unexpected circumstances.

LT1 (5): His basic grammatical structures and sentence patterns are consistently well controlled. Errors are found in some complex sentences and sometimes interfere with the meaning.

LT2 (4): Basic structures are not always well-controlled. He hardly used past tense when conversing upon the past events. Incomplete sentences are commonly found.

LT3 (4): Creatively uses basic grammatical structures and sentence patterns and usually well-controlled. Mixes up part tense and present tense a few times. Past forms of verbs are sometimes incorrect because of his pronunciation problem on –ed ending sounds. Mistakes on subject-verb agreement occur a few times which do not affect much of overall meaning.

LT4 (4): Though his grammatical structures and sentence patterns are used creatively but incorrect use of verb-tense forms is frequent. Occasionally, the verb of the sentence is omitted. However, he is, most of the time, able to get his messages across.

LT5 (5): - (LT5 did not make any specific comment. She just jotted down some errors that the test-taker made which were "attached with", "used to flew".)

LU1 (5): in operation, structures are controlled and do not interfere with meanings.

LU2 (4): He made quite a lot of mistakes in grammatical structure, but it seemed that this rarely interfered the meaning. He did quite well in the interview section.

LU3 (4): Errors can be often found when the test taker is in unusual situations. Grammatical structure used becomes more controlled when the test taker is in common communication situation like an interview.

LU4 (5): Some grammatical errors are observed but not interfere with meaning.

LU5 (4): Not many grammatical structures and sentences are produced creatively. Once getting the unexpected question, he sometimes seems to creatively create the grammatical sentences.

Vocabulary

Groups of Raters (Scores)	Remarks

OT1 (3): Again, good until abnormal situations. But he was able to let me know what he was experiencing. Also he was able to use the word "dynamite" when describing the bomb.

OT2 (4): Range of vocabulary is limited, but is able to paraphrase or use general terms.

OT3 (4): When talking about work-related topics, subject can communicate effectively. When talking about a non-familiar topic, he) is able to paraphrase successfully.

OT4 (4): Vocab. range is sufficient but need to paraphrase sometimes e.g. "the builder a/c", "give the pilot to fly". (However, it is) sufficient to communicate effectively.

OT5 (4): Range of vocabulary is sufficient to communicate on common work related topics. (He is) able to paraphrase when lacking vocabulary.

OU1 (5): Good vocab.

OU2 (5): Even using only plain words, it is also easy to understand and paraphrase or negotiate meaning is

not required.

OU3 (3): Still limited. He tends to use "something like that" quite often. Still have some trouble in finding the right word to describing things.

OU4 (5): Use only simple words to is understanding which is fair, but this will define the ability as "general".

OU5 (3): Sufficient under work related topics but limiting in lengthy descriptions.

LT1 (5): (He) has vocabulary range sufficient enough to effectively communicate on common and familiar topics including job-related matters. (He) can paraphrase successfully.

LT2 (4): A certain amount of vocabulary is enough for conforming effective communication in job-related matters. Even so, there are still some mistakes on word choice and collocations.

LT3 (4): Able to use simple vocabulary effectively and paraphrase successfully when lacking of them in unexpected situations.

LT4 (4): ID 1 has sufficient range of vocabulary to communicate effectively in work-related topics. However, when confronted with unexpected situations he has to grope for words (when he was trying to describe the foreign objects on the runway). The way he paraphrases the unusual situations on board is successful.

LT5 (6): Job related vocab. (was) used effectively. Appropriate wide range vocab.

LU 1 (5): Use concrete words and effectively relate to work.

LU2 (4): It seemed that he had quite a huge vocabulary range. This could be seen when he explained the existing situations. The problem was that the rater did not Have enough vocabulary range in this field, so the rater wasn't sure that the speaker made use of appropriate vocabulary or not.

LU3 (4): Vocabulary range and accuracy tends to be sufficient for common topic but in terms of unfamiliar topics, problem can be sometimes perceived.

LU4 (5): The pilot performs good control of the technical language especially used in ATC domain. Though, practice language to describe variety of objects would help improve language proficiency. *Should he just say "bomb" or "explosive device"?

LU5 (4): He satisfactorily produces wide range of vocabulary especially the vocabulary related to the work-related topic. He can sometimes paraphrase the terms that might not be familiar.

Fluency

Groups of Raters (Scores)	Remarks

OT1 (4): Link with structure and vocabulary, his fluency only slowed down in abnormal situations.

OT2 (5): Has good use of discourse markers and fillers to keep a smooth level of fluency.

OT3 (4): Usually when taking about a work-related topic, fluency is at an appropriate tempo. When talking about something spontaneously, some loss of fluency may occur but not distracting.

OT4 (5): (He is) able to speak with natural flow only small distracting fillers such as "em", "er" are shown.

OT5 (4): During normal conditions, he is able to use language at appropriate tempo. There is loss of fluency when faced with non-normal situations, but doesn't prevent effective communication.

OU1 (5): O.k.

OU2 (5): Well done during interviewing. (He) can speak at length and easy to understand.

OU3 (4): Acceptable and sufficient to get the point across in a timely manner.

OU4 (4): The major problem which reduces his fluency is the way he thinks in "Thai" prior to communicating. This is the general disadvantage of speakers of English who use it as a second or third language.

OU5 (4): Occasional loss of fluency during spontaneous interaction.

LT1 (5): He is a confident speaker and he can speak at length with relative ease on his familiar topics. (He) can make use of appropriate discourse markers or connectors. (He) can produce quite natural flow of speech generally.

LT2 (4): He is able to generate the language at length even though loss of fluency occurs from time to time when he has a difficulty searching for appropriate word choices. Discourse markers are rarely used.

LT3 (4): Speaks continuously with an appropriate tempo. Occasionally lacks flow only when trying to find appropriate words (paraphrasing). Able to use basic connectors in most cases, and fillers are not distracting.

LT4 (5): During face to face interview, his speech is spontaneous and he is at ease producing stretches of language as it is about his work and interest. There are a few occasions when his fluency is slowed, that was when he faced with difficult situations on board, but overall communication is not impeded.

LT5 (6): Quite fluent when speaking. No hesitation. Good speech flow. Speak quite naturally.

LU1 (4): Occasional loss of fluency when operating. The filler like "urr…" is heavily used during interviewing.

LU2 (3): He was fluent in the test part 2 and 3, but he paused many times in part 1. Also, he hesitated to answer questions.

LU3 (5): Generally fluent but the fluency speaking could be clearly seen when talking about common topic. Very few hesitations could be felt.

LU4 (5): Fluency is difficult to determine due to individual style but if compared with ATC staff, it is clear that the pilot should get training with respect to fluency.

LU5 (3): This test-taker seems to confront with the problem of fluency. Once producing the language, his communication still contains some chunks of words that might break down the communication.

Comprehension

Groups of	Remarks

Raters (Scores)	

OT1 (5): Very good comprehension. He seems to understand all of the situations.

OT2 (5): Good comprehension and ability to understand different situations, whether normal or non-normal (situation).

OT3 (4): For work-related topics, comprehension is mostly accurate. When some new topic arises, comprehension may slow down but he gets therein the end after some clarification.

OT4 (5): Mostly accurate and spontaneous.

OT5 (5): When speaker is confronted with situational complications, comprehension is slower than normal. Mostly accurate on work related topics.

LU1 (5): Comprehension in both operating (in situations) ad interviewing is accurate. Dialogues do not require clarification.

LU2 (4): It seemed that he understood what the interlocutor said, especially in part 2, 3. However, he paused many times in part 1. If it is the real emergency situation, it may cause problems. Also, the problem was the same as the vocabulary criterion. The rater also couldn't understand well in part 2. (I) need more background knowledge.

LU3 (4): Some comprehension problems can be found when the test taker was asked to paraphrase situations. However, in terms of work-related or common topics comprehension problem were hardly seen.

LU4 (5): Comprehension is all tasks have been indicated by task completion, although there were signs of "negotiation of meaning" such as asking for clarification.

LU5 (4): This test-taker can operate the comprehension concretely when discussing the work-related topic. When he confronts with the unexpected situation, he can moderately produce or respond with the situation.

Interactions

Groups of Raters (Scores)	Remarks

OT1 (4): Again, very good interactions for the interview.

OT2 (5): During the interview, the speaker is able to respond immediately and is able to give information concerning the subject at hand with ease.

OT3 (4): Interaction is good. (He) can handle most situations adequately.

OT4 (5): Responses are immediate, appropriate and interact with ease in nearly all situations.

OT5 (5): During normal situations, responses are usually immediate and appropriate. He is able to maintain dialogue with the interviewer and check, clarify and confirm what has been said when misunderstandings occur.

OU1 (6): Good interaction.

OU2 (4): Not quite well during answering the questions but did well during interview.

OU3 (4): Good interaction without shyness or reservation.

OU4 (4): Fair interactions between both speakers. The candidate shows slow responses when unexpected situations occur.

OU 5 (4): Usually immediate with no loss of continuity during unexpected events.

LT1 (6): He interacts with ease most of the time. Usually responds appropriately and immediately. (He) can manage effective relationship with the interviewer.

LT2 (4): Responses are generally immediate and informative. The exchanges are pretty natural.

LT3 (5): Able to respond immediately, appropriately, and informative. Relaxed and manage speaker/listener relationship effectively.

LT4 (5): His responses are mostly appropriate and informative. The conversations are spontaneously maintained throughout the interview.

LT5 (6): (He) could perform immediate responses appropriately. Answer the questions appropriately.

LU1 (5): Give immediate (and) appropriate responses. (He) could deal with expected and unexpected turns effectively.

LU2 (4): Overall, he could interact well, but sometimes it wasn't interactive. He did very well in part 3.

LU3 (5): The test taker can almost always give immediate response to the prompts.

LU4 (5): Interactions, slowness in response was observed occasionally, especially during communications with ATC but during the interview there was no sign of slow response.

LU5 (4): This man can interact in the communication in an operational level. Once confronted with the unexpected event, he can adequately clarify and interact.

General/Overall

Groups of Raters (Scores)	Remarks

OT1 (4): After facing abnormal situations the subject starts to show errors in sentence structures, vocabulary. Thus his fluency also suffers. However I was able to understand his intentions.

OT2 (4): Thai accent is present but understandable. More emphasis on structure and vocabulary will be beneficial for an increase in level.

OT3 (4): Standard. (There is) nothing more, nothing less.
OT4 (4): Interviewee is confident and manages to interact very well.

OT5 (4): Normal ATC phraseologies are standard, but when faced with non-normal situations, fluency, structure and interaction decrease slightly.

OU1 (5): Overall is good. (There are) some errors in grammar and pronunciation.

OU2 (5): Generally fair during answering the questions. (His) disadvantage was how to create a sentence and time consumed to react for non work-related questions but well done on interview.

OU3 (4): Although somewhat limited in terms of vocabulary and grammar, his clear and articulated speech style get the point across.

OU4 (4.5): Able to communicate in English with some difficulties as his major language structures. (He is) fair in English command which is understandable by all users.

OU5 (4): Operational level of English proficiency. (It is) sufficient for a safe conduct of flight.

LT1 (5): A confident speaker who can handle both general English and work related English quite well.

LT2 (4): He needs to extend the vocabulary range and strictly monitors grammatical structures used as well as regularly practices on the pronunciation skill.

LT3 (4): To aim for a higher level of language proficiency interview, please practice continuously on weak areas mentioned on the attached form.

LT4 (4): ID 1 can achieve Level 5 easily if he pays more attention to his pronunciation and structure which are already at a "high 4". (He) just needs to speak English more often to recall some needed words to express himself.

LT5 (5): Some small problems with final sounds, port of speech but can be understood easily. Communication (is) quite fluent. Good comprehension and interactions.

LU1 (5): It"s satisfactory for overall.

LU2 (4): (LU2 did not make any remark on this item.)

LU3 (4+): (LU3 did not make any remark on this item.)

LU4 (5): The pilot performs all the tasks required as its intended purposes. However, language proficiency in terms of word pronunciation, grammatical structure and fluency can be improved to attain higher level of language proficiency.

LU5 (4): The test-taker proficiency level in overall is in the operational level. He can moderately communicate the gist information with appropriate non-verbal language, like intonation. However, his communication fluency still contains a high level of chunks.

Appendix G

Raters‟ remarks on speech sample number 2

Pronunciation

Groups of Raters (Scores)	Remarks

OT1 (3): Subject has strong Thai influence in his pronunciation. However I was able to understand his intentions.

OT2 (4): A heavy Thai accent is present. Some words are not easily understandable. (There are) problems with final sounds, clusters, stress and intonation.

OT3 (3): Pronunciation (is) heavily influenced by first language, frequently interferes with ease of understanding.

OT4 (4): Problem with "r" and "l" e.g. "ellor" instead of "error", "aelodome" instead of "aerodrome", "alival" instead of "arrival", "leport" instead of "report", "lunway" instead of "runway". Speak without final sound e.g. "hell" instead of "help", "decimon" instead of "decimal", "devide" instead of "device". Leave "l" and "r" sounds in words e.g. "tire bow" in "tire blow", "cimbling" in "climbing", "form" in "from", "cock" in "clock", "expotion" in "explosion", "explosive" in "explosive". Pronunciation and intonation are influenced by the first language but still understandable.

OT5 (3): Heavily influenced by mother language and usually interfere with ease of understanding. "R" and "l" sounds misuse.

OU1 (5): Some words are wrongly pronounced but understandable e.g. "real flight" (is) pronounced as "leal flight".

OU2 (5): Generally above level 4 as not a heavy accent but the disaster will come from the rest.

OU3 (2): Inarticulate and still has a strong Thai accent.

OU4 (4): Fair pronunciation which only familiar listeners who get used to will understand.

OU5 (3): Influenced by first language with frequent interference with ease of understanding.

LT1 (4): He has quite strong Thai interference and he speaks with many problems of r/l substitution, final sound omission and incorrect stress/intonation. Sometimes it interferes with the ease of understanding.

LT2 (3): Rather strong Thai interference with numbers of errors on all area. Sometimes it is rather difficult to understand some of the words he produces.

LT3 (3): His overall pronunciation is difficult to understand due to many problems. For example, dropping final sounds (f, v, l, s, ge, x, ft), substitution of r and l sounds, dropping clusters with r and l, distorted vowel sounds, and –s and –ed ending sounds.

LT4 (4): Like most Thai speakers of English, his pronunciation is strongly interfered by his L1. There are some problems of sound distinction and cluster sound production. Some words are intelligible but still he can make himself understood as the word stress and intonation patterns are used effectively.

LT5 (4): - (LT5 did not make any specific remarks but she jotted down some pronunciation errors such as no final sounds e.g. „reques‟ for „request‟, „aircraf‟ for „aircraft‟, „lef‟ for „left‟, „dewide‟ for „device‟, „decimon‟ for „decimal‟, or adding „r‟ sound where there was no „r‟ e.g. trow‟ for „tow‟, etc.)

LU1 (3): The insufficient of stress and intonation obviously interfere (with) understanding. It‟s rather hard to understand some words.

LU2 (3): His pronunciation is influenced by the first language. He couldn‟t speak clearly in some sentences. This might be because he wasn‟t confident.

LU3 (3): Pronunciation problems could be perceived throughout the test taking. Many times it was hard to understand what the test taker was saying.

LU4 (4): Ambiguity (is) observed with many word pronunciations.

LU5 (3): The pronunciation, stress, rhythm and intonation are mostly influenced by the first language, Thai. This can interfere with the ease of understanding.

Structures

Groups of Raters (Scores)	Remarks

OT1 (3): During standard operations there are sentence structure and grammatical errors.

OT2 (4): Tenses are sometimes used wrong. Minor grammatical errors occur, especially in unexpected situations.

OT3 (3): Sentence structure (is) not very well controlled. Errors occur frequently and interfere with meaning.

OT4 (4): (Incorrect) word form e.g. “turbulence moderate” instead of “moderate turbulence”. Verb to be (are wrongly used). Sometimes start sentences without subjects. Errors occurred but still understandable.

OT5 (4): Basic sentences are usually well controlled. Errors occur in unusual or unexpected situations and sometimes interfere with meaning.

OU1 (4): Fair. (Some mistakes such as) “it‟s not a seldom, it‟s not problem, all that you talk about”.

OU2 (4): Even poor English structure but still show their genuine meaning.

OU3 (3): Grammar is not bad and can complete sentences well enough to understand.

OU4 (3): All sentence structures are based on his own language structures which lead to misunderstanding or totally no understanding.

OU5 (3): Grammar & structures (are) not well controlled and frequently interfere with understanding.

LT1 (4): He sometimes uses incomplete sentences and makes many grammar mistakes especially in unusual or unexpected situations. He is still understandable most of the time.

LT2 (3): He is unable to consistently control even the basic grammatical patterns. Short phrases and incomplete sentences are generally generated. Past tense construction is rarely applied even when he narrates the past events.

LT (3): Creates short basic grammatical structures and sentence patterns with inconsistent control of past tenses, parts of speech, subject-verb agreement. In some unexpected situations, incomplete sentences are found.

LT4 (4): Most of his sentences are short and simple and are well controlled. However, minor grammatical errors are frequent but only sometimes affect ease of understanding.

LT5 (3): Quite clear information delivery e.g. „explosive". (LT5 also jotted down some grammatical errors e.g. „everything have got a error", „we do that in a parallel".)

LU1 (4): Basic grammatical structures are used in real time conversation.

LU2 (4): Overall was fine.

LU3 (3): Many structural mistakes. However, what the test taker intended to deliver was generally understandable.

LU4 (4): Because phraseologies are used as normal procedure, structures beyond fixed phrases are problems for pilots.

LU5 (3): The consistency of the grammatical and sentence patterns is not well enough monitored. This generates the errors that can interfere the overall understanding.

Vocabulary

Groups of Raters (Scores)	Remarks

OT1 (3): In part 2 - the FOD situation, the subject did not understand what ground control asked him.

OT2 (4): Limited vocabulary, but able to paraphrase. The speaker has the ability to use some idioms on certain occasions.

OT3 (4): When talking about work-related subjects, vocabulary (is) adequate but when talking about something unfamiliar vocabulary is lacking but (he) can successfully paraphrase.

OT4 (4): Small choice of vocabulary but sufficient to communicate. Paraphrasing is not always successful e.g. "we got fly 747-400". (There is) limited use of vocabulary.

OT5 (4): Vocabulary dealing with work related topics are sufficient. (He is) able to paraphrase in unexpected situations.

OU1 (4): (He) cannot use relevant vocabulary but can explain e.g. "bomb blasting on cabin make me looking".

OU2 (4): Only for plain & easy words is enough. If try harder, the situation will go wild.

OU3 (3): Limited. Use (the) phrase "something like that" often when cannot find the appropriate word.

OU4 (3): Limited in range and accuracy of vocabulary which are insufficient to communicate in normal situations.

OU5 (3): Sufficient on work-related topics but limited & sometimes inappropriate otherwise.

LT1 (4): His vocabulary range is often sufficient to communicate effectively on common familiar topics.

(He) can paraphrase successfully when lacking words in unexpected situations.

LT2 (3): The range of vocabulary is rather limited to converse in unfamiliar matters. Inappropriate word choices are commonly found in the speech.

LT3 (3): Vocabulary range and accuracy are often sufficient to communicate on common, concrete or work-related topics but often selects inappropriate word choices. Also, unable to paraphrase successfully because limited range of vocabulary.

LT4 (4): His vocabulary stock is rather limited but enough to talk about his work and related topics. When prompted, he is able to paraphrase but within his area of interest (experiences and flight procedure).

LT5 (3): Job related vocab. Quite limited vocab. Difficult to find an appropriate word to explain the material on the runway. Sometimes (he) used unclear vocab.

LU1 (4): Words related to work are appropriately used. However, they are quite limited in spontaneous speaking, especially when the person elaborates the interviewer‟s questions regarding the advantages and disadvantages of technology in aviation.

LU2 (4): It sounded like he had quite a huge vocabulary, but he still couldn‟t explain well in some situations.

LU3 (2): The test taker tended to have difficulty in expressing thoughts on describing things or situations.

LU4 (4): The pilot should (have) good control of technical language used in this specific context.

LU5 (3): The test-taker adequately acquires the vocabulary that relates to his work. But he sometimes uses the word choices that are not unable to paraphrase into the ease of understanding.

Fluency

Groups of Raters (Scores)	Remarks

OT1 (3): Overall (is) good but when pushed, his fluency drops.

OT2 (4): Has fluency on an operational level, but sometimes hesitates when unexpected problems/situations occur. More use of gap fillers would help.

OT3 (3): Starts and stops in inappropriate spurts. Pauses and slowness are distracting and prevent effective communication.

OT4 (4): Minimal use (of) stylistic effect but the tempo is o.k. Limited use of connectors between sentence.

OT5 (4): Occasional loss of fluency on spontaneous interactions or unexpected events.

OU1 (5): O.k. but sometimes got stuck due to limited vocabulary.
OU2 (4): (He) can produce some stretches of language.

OU3 (2): Have trouble with descriptions and still speak with limitation of aviation phraseology.

OU4 (4): Rather fair in English skills which reduce his confidence, ability and fluency in command and control of his English communication.

OU5 (2): Frequent hesitations. (There are) mainly short memorized utterances.

LT1 (4): (He) can produce stretches of language to express himself. Occasionally loss of fluency but he can move on and have an effective communication. (He) can make limited use of connectors and discourse markers.

LT2 (3): The fluency is dropped when he encounters the language difficulties. Fillers are commonly found and sometimes are quite distracting.

LT3 (4): Produces stretches of language but with rather fast speech ate. In some unfamiliar situations, there may be occasional loss of fluency. Also make limited use of basic connectors.

LT4 (4): He uses some discourse markers to buy time when he is confronted with language difficulty. He can also speak at length when prompted. However, many of his responses connecting with his work are formulaic and memorized.

LT5 (4): Not fluent. Not good when answering unexpected questions. Not very good tempo.

LU1 (4): The person attempts to carry the conversation even sometimes clarification is needed.

LU2 (4): He was quite fluent, but sometimes he wasn‟t sure what he should answer. Therefore, he paused a few time.

LU3 (2): Hesitation was often shown. The delivery of thoughts or opinion was not sooth and sometimes the test taker was unable to respond.

LU4 (4): Should be improved by more practicing and training.

LU5 (3): The test-taker seems to have a problem of pausing. He usually produces the stretch of sentence without any appropriate pausing. His slowness in fluency can undermine his productive skill.

Comprehension

Groups of Raters (Scores)	Remarks

OT1 (4): Most parts he understands the situations he was in.

OT2 (3): Speaker requires detailed clarification to understand some questions. A misunderstood question resulted in a completely wrong answer.

OT3 (3): Not so much of a problem when talking about work-related subjects but when confronted by an unexpected event, (he) has great difficulty understanding.

OT4 (4): Misunderstand some of the questions but can be clarified. (He) could be more spontaneous in responses.

OT5 (4): Comprehension on common work related topics is sufficient. When faced with unexpected turn of events, comprehension is slower.

OU1 (5): In section 4, (the) interviewer had to repeat the question again but finally conversation was understood.

OU2 (3): For non-normal situation, an outcome is a different story that may cause confusion or misunderstanding.

OU3 (3): Have trouble in understanding the examiner‟s more complex questions. (He) seems to reply quickly and prematurely after the questions were asked. (He) should process the questions more.

OU4 (3): Limited to familiarized words or phrases which reduce the ability to get the messages through.

OU5 (3): Frequent misunderstandings on non work-related topics.

LT1 (4): Comprehension is mostly correct but when he confronts with a linguistic or situational complication or unexpected turn of events, he becomes slower or needs to be clarified.

LT2 (3): He has difficulties in answering some of the questions which requires repeated classification from the interviewer. Sometimes he still provides irrelevant replies even though the interviewer tries very hard to guide him on to the right track.

LT3 (4): Comprehension is mostly accurate on general and job-related topics but when confronting with difficult or unexpected situations, comprehension may be slower and need clarification strategies.

LT4 (4): He occasionally needs the repetition of the questions but most of the time his comprehension is accurate.

LT5 (3): Some questions needed to be asked twice to make the interviewee understand and to answer very easily. Not very good comprehension.

LU1 (3): Repetition is needed sometimes in the conversation. The person also fails to understand the situation and cannot elaborate his answer well e.g. failed to describe the cause of explosive device.

LU2 (4): He could answer what the interlocutor asked, but sometimes he needed (the interlocutor) to repeat the questions.

LU3 (3): Comprehension problems could be perceived in all parts of the test. Many times the test taker didn‟t understand the interviewer‟s questions, though the questions were not complicated. Questions often needed simplification.

LU4 (4): More listening practice needed.

LU5 (3): Comprehension is moderately on common when the topic is related to his work. Sometimes his problem on the communicative skill probably originates from his language proficiency, not from his linguistic proficiency.

Interactions

Groups of Raters (Scores)	Remarks

OT1 (4): Good for the interview.

OT2 (4): Immediate responses were given. Speaker did confirm with interviewer when not sure.

OT3 (3): Adequate when talking about work-related topics but stutters a lot when talking about something unfamiliar.

OT4 (4): Responses are immediate and informative.

OT5 (4): Responses are usually immediate and can handle unexpected events and non-normal situations. (He is) able to check and confirm to clarify correct understanding.

OU1 (5): Good.

OU2 (4): Doubtful during answering the question but recover during interview by confirming and clarification.

OU3 (2): Not comfortable in communicating with the examiner and find it hard to express his point of view. ? it in his conversation.

OU4 (3): Limited ability to interact with other English speakers. The tester indicates his usual misunderstanding when the interviewee expressed his main ideas which lead to miscommunication.

OU5 (3): Sufficiently responsive on familiar topics but inadequate on others.

LT1 (4): His responses are usually appropriate and immediate. (He) can maintain exchanges even when dealing with unusual situations. (He) asks for clarification when misunderstanding occurs.

LT2 (3): Responses especially upon complicated matters are not quite immediate and informative. The exchanges are not thoroughly smooth because of the misunderstanding causing a short silence from time to time.

LT3 (4): Usually responds with some information and able to ask, check, and confirm when misunderstanding occurs.

LT4 (5): This is his strongest area of all the six areas being checked. He deals adequately with misunderstanding by checking, confirming and clarifying.

LT5 (4): Quite accurate, appropriate. Quite immediate responses.

LU1 (4): The person gives immediate responses rather well, but sometimes the responses are not informative. It needs more elaboration.

LU2 (4): He showed that he could interact immediately, but sometimes it was not informative.

LU3 (2): The test taker was often unable to give immediate responses.

LU4 (4): Repetition is normal procedure for language use in this context. The pilot knows the drill of turn-taking, asking for clarification.

LU5 (4): His response in the communication is informative. He can initiate or maintain the conversation. He sometimes shows his confidence to exchange or deal with unexpected situation.

General/Overall

Groups of Raters (Scores)	Remarks

OT1 (3): (In) most standard operations the subject did well but there were problems with grammar and vocabulary errors. Thus fluency suffers.

OT2 (3): Strong Thai accent, continuous problems with final sounds, clusters and stress but does not affect meaning. Additional emphasis on pronunciation, structure and vocabulary will be beneficial to the speaker.

OT3 (3): - (OT3 does not make any remark on this item.)

OT4 (4): (He) should take pronunciation class.

OT5 (3): Pronunciation is heavily influenced by mother tongue and interferes with the meaning of words and sentences. Other criteria are fine to operate under normal situations.

OU1 (5): Some errors in structure and vocabulary. Limited vocabulary makes conversation, sometimes, interrupted but interviewee can explain and make conversation go through.

OU2 (4): Only good on work related topic and may have some difficulty during unexpected or complicated situation.

OU3 (3): Need to be more articulate. Listen before answer. Basic phraseology still (is) not standard. (It is) vague in expressing himself.

OU4 (3.3): Limited English communicative skills which may lead to misunderstanding, even in normal situations. His general English is only at or below the standard requirements.

OU5 (3): Below operational level.

LT1 (4): - (LT1 did not make any remark on this item.)

LT2 (3): He needs to enhance the language in every aspect in order to perform effective communication.

LT3 (3): Weak areas on his pronunciation should be paid attention to first as his level could drop to "2" if more mispronunciation is found.

LT4 (4): ID2 is at a borderline „Level 4". Any questions outside his area of familiarity would cause problems for him and might affect his „level".

LT5 (3): Problems with pronunciation (final sounds, clusters). Some sentences (are) not very clear (sometimes) when speaking. Comprehension (and) fluency (are) not very good.

LU1 (4): It"s generally comprehensible.

LU2 (4): - (LU2 does not make any remark on this item.)

LU3 (3): - (LU3 does not make any remark on this item.)

LU4 (4): Although this pilot could complete tasks required in the simulation test, he should get more training on language proficiency in all aspects. Because language used in this context seems to be quite fixed, he could do better.

LU5 (3.5): The test-taker proficiency level in overall is in between pre-operational and operational level. His pronunciation seems to lower his comprehension and productive skill.

Appendix H

Raters‟ remarks on speech sample number 3

Pronunciation

Groups of Raters (Scores)	Remarks

OT1 (4): Subject had some first language influences in his pronunciation but he has good rhythm.

OT2 (4): Problems with cluster sounds and stresses are apparent. Final sounds are consistent with Thai accent.

OT3 (5): Even though pronunciation is influenced by the first language, it rarely interferes with ease of understanding.

OT4 (4): Problem with "r" and "l" sounds e.g. "advisoly" instead of "advisory", "leal" instead of "real", "lequest" instead of "request". Pronunciation and intonation are heavily influenced by the first language.

OT5 (4): Pronunciation, stress and rhythm are somewhat influenced by mother tongue language but only sometimes interfere with ease of understanding.

OU1 (4): Fair. Sometimes not understandable with that word alone e.g. "put back, start up", "hone shot runway 26L", "runway two sick left".

OU2 (5): Mostly influenced by the first language but still easy to understanding.

OU3 (4): Good, clear and understandable. He has patience to explain himself in an articulate way.

OU4 (5): Strong Thai accent which has no limitation in understanding his English even to a fair listener.

OU5 (4): Some first language influence with little interference with ease of understanding

LT1 (5): His pronunciation, stress and intonation are influenced by Thai language but rarely interfere with the ease of understanding.

LT2 (4): Rather strong Thai interference mostly on stress and intonation patterns. At word level, there are many errors occurred such as final sound/cluster droppings and r/l substitutions but it is still understandable.

LT3 (3): Able to make a clear speech in the beginning but unable to consistently control pronunciation errors later on which often interfere with ease of understanding. Errors are e.g. r/l substitution, dropping of clusters with r and l, non-shared sounds (th, v, j), and dropping of cluster final sounds (ft, st).

LT4 (3): ID 3‟s pronunciation is acceptable during the R/T communication, but he was very difficult to understand during the interview as many of his words are mispronounced. His pronunciation is heavily influenced by his L1 and often causes misunderstanding.

LT5 (4): - Not quite good intonation, stress. (LT5 also jotted down some pronunciation errors such as no or wrong final sounds e.g. „reques‟ for „request‟, „firs‟ for „first‟, „aircraf‟ for „aircraft‟, „advantate‟ for „advantage‟, or no „l‟ sound e.g. „fight‟ for „flight‟.)

LU1 (5): The stress and intonation are used appropriately and effectively. Generally, there is no difficulty in understanding.

LU2 (3): Actually, he did well at the beginning, but after pausing and hesitating many times he couldn't control his rhythm.

LU3 (4): Pronunciation is generally understandable, not usually interfere meaning.

LU4 (4): LU4 did not make any remark on this item.

LU5 (4): His pronunciation, stress, rhythm, and intonation are not much interfered by the first language. However, his receptive skill sometimes interferes with the ease of understanding.

Structure

Groups of Raters (Scores)	Remarks

OT1 (3): During part 4, there were some structure and grammatical errors.

OT2 (3): Sentence structure and patterns are sometimes broken. Grammatical errors do occur but meaning is intact.

OT3 (4): Grammar and sentence structure are well controlled and although errors do happen, they do not interfere with understanding too much.

OT4 (3): Gerund form mixed with verb to be e.g. "make me does not looking a/c in", "may be make a pilot freeze", "due to we have malfunction". (Incorrect) word form e.g. "the future can improve", "do himself for follow …". (Wrong use of) verb to be e.g. "Device over the bin" – it should have "is", "make we safe", "pilot must awareness".

OT5 (4): Errors occur in unusual or unexpected events or situations but rarely interfere with meaning. Basic sentences are well controlled.

OU1 (4): Fair. Errors in grammar.

OU2 (4): Have some errors in unusual situation, even using basic grammatical structures but not interfere with meaning.

OU3 (2): Still weak in forming complete sentences.

OU4 (4): Few grammatical errors in unfamiliar situations. The sentence structure is fair to understand his expression.

OU5 (4): Some errors with unusual circumstances but mainly well controlled.

LT1 (3): His basic grammatical structures and sentence patterns used in predictable situations are not always well controlled. He usually uses only key words and drops function words or makes mistakes on tenses, part of speech and other grammar elements. Mistakes often interfere with meaning.

LT2 (3): He is rather lack of language foundation. Incomplete sentences and short phrases are generally generated. There are countless errors which strongly interfere with the meaning.

LT3 (3): Basic sentence structures are not always well-controlled, and often translated from Thai. In unexpected circumstances, he creates incomplete sentences.

LT4 (3): He speaks in phrases and sometimes only words are used. Some of his grammatical structures and sentence patterns are memorized or rehearsed.

LT5 (4): - (LT5 did not make clear remark. She just wrote „may be excited", „article the")

LU1 (5): Structures are well controlled in operation.

LU2 (4): He made some mistakes, but (I) still understood what he would like to explain.

LU3 (3): Structural problems could be seen throughout the test especially in the second part when the test taker was required to speak at length.

LU4 (4): - (LU4 did not make any remark on this item.)

LU5 (4): He can generate wider range of grammatical structures and sentences creatively. However, the errors sometimes occur when he has to generate the grammatical sentence in some circumstances.

Vocabulary

Groups of Raters (Scores)	Remarks

OT1 (3): Vocabulary range is limited. In part 3 when describing the bomb, he did a good job, but in part 4 (future technologies) he had problems expressing his thoughts.

OT2 (3): Range of vocabulary makes the speaker limited to certain words. Sometimes paraphrasing is difficult.

OT3 (4): Most of the vocabulary range is sufficient. (He is) able to paraphrase successfully when talking about non-familiar topics.

OT4 (4): Limited choice of vocabulary e.g. "not follow the a/c computer", "we got to do a/c told us" – it could be "according to the instruments" but (he) can paraphrase when lacking vocabulary.

OT5 (4): Vocabulary range is sufficient to communicate on work related topics. (He is) able to paraphrase when lacking vocabulary in unexpected situations.

OU1 (4): Fair. Limited vocabulary.

OU2 (4): Using only plain words is sufficient.

OU3 (2): Limited, hence making it hard for him to speak outside of aviation phraseology or aviation vocab.

OU4 (5): Able to communicate effectively as his variety of vocabulary selection.

OU5 (3): Limiting in non work related circumstances.

LT1 (3): (He) has adequate vocabulary range to communicate on common familiar matters. Sometimes he pauses and groups for words. (He) can't paraphrase successfully or uses wrong word choice.

LT2 (3): He has got rather limited amount of vocabulary. Several times he is unable to either search for proper word choices or use them accurately. The ability of paraphrasing is out of question.

LT3 (3): vocabulary stock is limited especially when communicating on some familiar job-related topics and also unexpected ones. (He is) unable to paraphrase successfully.

LT4 (3): Word choice is ID 3"s problem and it shows his limited vocabulary range. Describing and paraphrasing are almost impossible.

LT5 (4): Job related vocab. (was) used appropriately

LU1 (5): Words related to work are used effectively.

LU2 (3): He couldn"t explain well. This might be because he didn"t have enough vocabulary knowledge.

LU3 (3): There is sufficient vocabulary range in work-related. But in common communication, it seems the test taker had difficulty to find appropriate words to express.

LU4 (4): LU4 did not make any remark on this item.

LU5 (4): He acquires a wide range of vocabulary and is able to paraphrase more successfully when confronting with unexpected circumstances.

Fluency

Groups of Raters (Scores)	Remarks

OT1 (4): (The candidate) suffers when he is describing abnormal situations. Fillers are sometimes distracting.

OT2 (4): Speech is fluent and at an understandable tempo. Fluency tends to drop when encountered by an unusual situation.

OT3 (4): Mostly the conversation was conducted at an appropriate tempo. Sometimes when transitioning to an unfamiliar topic, there was a loss of fluency but this was not distracting.

OT4 (4): Interviewee produced an appropriate tempo but occasional loss of fluency when asked about TCAS experience and new technology in a/c. Lots of "er", "ah".

OT5 (4): Loss of fluency in unexpected situations but can still effectively communicate and maintain effective communication.

OU1 (5): Fair. Not fluent due to grammar and vocabulary.

OU2 (4): (He) can produce a stretch of language with some difficulty.

OU3 (3): Still need to think before he speaks, making him appear slow but good in getting points across.

OU4 (5): Still demonstrated some difficulties as the test taker performed his thought in Thai prior to translating to English. This is very common to any user of English as a second language.

OU5 (3): Many pauses & consists mainly of short phrases.

LT1 (4): Although he can produce stretches of language, there are occasional losses of fluency when he hesitates or looks for words. He can make limited use of connectors and discourse markers.

LT2 (3): Short silence occurs from time to time when he has difficulties finding the words and constructing the language. Fillers sometimes are distracting.

LT3 (3): Produces stretches of language but inappropriate pausing is often found. Sometimes needs time to process the language, and fillers are sometimes distracting.

LT4 (3): There are hesitations in processing the language as his structures and vocabulary are not well established, and this, most of the time, impedes effective communication.

LT5 (4): Not very fluent. Not very good tempo. Not quite smooth when speaking.

LU1 (5): The speech flows spontaneously.

LU2 (3): He paused many times and also spoke slowly when he hesitated.

LU3 (3): Hesitation could be often seen. The delivery of responses was not smooth, especially in the interview.

LU4 (4): LU4 did not make any remark on this item.

LU5 (4): He can produce the whole stretch of complete sentences. However, he sometimes lost his fluency because of the formulaic speech.

Comprehension

Groups of Raters (Scores)	Remarks

OT1 (4): Good. He understands most of the situations he was in.

OT2 (4): Speaker has a good comprehension of the questions asked and can clarify when unsure.

OT3 (4): When talking about work-related topics, it went very well. (He) was a little slower when talking about something unfamiliar but generally more than acceptable.

OT4 (4): (He) is consistently accurate in nearly all contexts including a limited range of speech varieties and vocabulary.

OT5 (5): (He is) able to comprehend on work related topics when confronted with unexpected turn of events.

OU1 (4): Sometimes questions need to be repeated.

OU2 (4): Mostly accurate by using easy and plain words.

OU3 (3): Still having some trouble understanding the examiner‟s words & phrases. And (he) had some confusion with regard to certain questions.

OU4 (4): Comprehension is fair, but get slower in some complex situations.

OU5 (3): A lot of misunderstanding on non routine communication.

LT1 (4): Mostly accurate comprehension on common familiar topics but needs to be clarified when facing difficulties – linguistically or situationally.

LT2 (4): He is able to comprehend most of the questions. Still, there are several times that he needs to ask for clarification.

LT3 (3): Comprehension is often accurate on general and familiar job-related topics, but may fail to understand different accents or linguistic or situational complication.

LT4 (3): When facing with unexpected turn of events, ID 3 has problems understanding the situation and needs a lot of guiding questions before he can come up with the responses. However, comprehension is accurate only in the area of routine work.

LT5 (5): Sometimes (he was) not sure when answering questions. Few questions (must be) asked twice.

LU1 (5): The person gives good description related to situation given.

LU2 (4): It seemed that he understood questions, but it was difficult for him to explain. Also, he couldn‟t comprehend some sentences.

LU3 (4): Comprehension of work-related topic is generally correct.

LU4 (4): LU4 did not make any remark on this item.

LU5 (3): He seems to have problem on the receptive skill rather than productive skill because he cannot receive or get the gist information. However, his comprehension to produce language is much better.

Interactions

Groups of Raters (Scores)	Remarks

OT1 (4): Good for the interview.

OT2 (4): Speaker responds in a timely manner and can maintain an appropriate conversion.

OT3 (5): (He) did very well in the speaker/listener interactions. (He) was good at carrying on with the conversation.

OT4 (4): Responses are mostly immediate and appropriate. Some misunderstanding was clarified by checking and confirming.

OT5 (4): Responses are usually immediate and checks, clarifies when not sure of questions when dealing with unexpected events.

OU1 (4): O.k. but with a lot of interruption in conversation.

OU2 (4): Require more immediate response if in doubt.

OU3 (3): Easy to understand ever though structurally weak in sentence forming. Appear willing to initiate conversation and ask questions when feel unsure.

OU4 (5): Good interaction with others in most situations. Other means of communication lead to his effective English skill.

OU5 (4): Mainly immediate & appropriate.

LT1 (4): His responses are usually immediate and appropriate. (He) can deal adequately with misunderstandings by checking or asking for clarification.

LT2 (4): He is able to maintain the exchanges quite well. However, the responses are not always appropriate and informative. He always needs times to think especially when he has to deal with more complicated questions.

LT3 (3): Responses are often immediate but not informative and not appropriate. Asks, checks, and confirms with inappropriate language patterns.

LT4 (4): ID 3 can maintain routine exchanges and deal adequately with misunderstanding as he is able to check, confirm and clarify the problems.

LT5 (5): Quite appropriate responding. (He) could perform immediate responses sometimes.

LU1 (5): Response time is very immediate.

LU2 (4): He wasn‟t sure sometimes, so he still needed clarification and confirmation.

LU3 (4): The test taker can generally give immediate responses especially in the first section. In the second section, the test taker was quite reluctant.

LU4 (4): - (LU4 did not make any remark on this item.)

LU5 (3): His response is sometimes appropriate. As said, he can generate the language productively but he inadequately deals with unexpected turn of events.

General/ Overall

Groups of Raters (Scores)	Remarks

OT1 (3): The subject did well. He had some problems with structure and vocabulary but he tried to use other methods to get his point across.

OT2 (3): Grammatical errors and limited vocabulary. Cluster sounds are more prevalent. Sentence structure needs adjustment.

OT3 (4): - (OT3 does not make any remark on this item.)

OT4 (3): Interviewee pronunciation is rather influenced by the first language. His deficiency in grammar and structure explains why he was rated overall in level 3.

OT5 (4): (He is) able to communicate at an operational level in all six criteria.

OU1 (4): Pronunciation is not clear with errors. Some wrong grammar. Limited vocabulary. Communication (conversation) needs quite many explanations to be understood.

OU2 (4): (He) has some difficulty to create a sentence but does well by using easy and plain words.

OU3 (3): Cool, slow and articulate (which are the) qualities that help to offset his weakness in comprehension and self-expression.

OU4 (4.75): Able to express ideas and make effective communication with any English users.

OU5 (4): Marginal operational level of English proficiency.

LT1 (3): - (LT1 did not make any remark on this item.)

LT2 (3): He needs to practice more on pronunciation and try to expand vocabulary range as well as control the grammatical patterns when generating the language.

LT3 (3): Practice hard on every area could contribute to his achievement on level 4 of Language proficiency interview test.

LT4 (3): It is interesting that this test taker can pronounce English words clearly and correctly in R/T communication. But for everyday English, he is hardly understood. (He) needs a lot of improvement in five out of six areas above.

LT5 (4): Quite good pronunciation, easy to understand. Grammatical structure needs to be improved. Vocab. range needs to be increased.

LU1 (5): It‟s satisfactory for overall.

LU2 (4): - (LU2 did not make any remark on this item.)

LU3 (4): **-** (LU3 did not make any remark on this item.)

LU4 (4): Overall language proficiency of this pilot sufficiently serves all tasks required in maneuvering/flying his aircraft and communicating with ATC. However, language training would surely give him more confidence in language communication and proficiency. *Language training on describing objects is needed.

LU5 (4): The test-taker proficiency level in overall is in the operational level. He can operationally communicate. However, he still lacks of the comprehension skill to interact in some unexpected situations.

Biography

Sutas Dejkunjorn, the researcher, is a Thai male, 53 years of age. He graduated with his first Bachelor‟s degree in Biology from the Faculty of Science, Chiangmai University in 1979. Twenty years later, he received another Bachelor‟s degree in Business Administration from Sukhothai Thammathirat Open University. After that, he earned his Master‟s degree in English for Specific Purposes from the Faculty of Humanities, Kasetsart University in 2005. He is a full-time employee working for Thai Airways International PLC (THAI). He has been flying as a pilot for THAI since 1981. Sutas was an MD-11 simulator instructor and flight instructor. He was also a Boeing 747-400 supervisory pilot and acted as Head of the Pilot English Language Development and Assessment Working Group. Sutas is presently a 747-400 captain and acting, unofficially, as Chief of Pilot English Language Assessment Group for THAI. He also teaches Aviation English as a guest lecturer for THAI Flight Crew Language Training Department. Sutas has his special interest in English usage in aviation context. Another area of interest is concerned pilot English proficiency assessment and evaluation, which inspired him to study in this field.