LANCASTER
UNIVERSITY
Department of Linguistics
and English Language

DEPARTMENT OF LINGUISTICS AND ENGLISH LANGUAGE

LANCASTER UNIVERSITY

What do ICAO Language Proficiency Test Developers and Raters Have to Say about the ICAO Language Proficiency Requirements 12 Years after their Publication?

A qualitative study exploring experienced professionals’ opinions

Angela Carolina de Moraes Garcia

Dissertation submitted in partial fulfillment of the requirements for the MA in Language Testing (by distance) degree

(19,652 words)

October 2015

Abstract

The International Civil Aviation Organization (ICAO) standard and recommended practices (SARPs) related to the language use for aeronautical radiotelephony communications were published in March 2003. Twelve years after their publication, in the light of research suggesting the revision of the ICAO policy, it is important to learn what experts who have been working with the ICAO language proficiency requirements (LPRs) think are their strengths and weaknesses according to their experiences. This dissertation investigates experienced test raters and test developers' opinions about the ICAO LPRs. Six expert professionals were interviewed in this qualitative analytic research and the data were analysed in accordance with the thematic analyses method. The discussions included not only general features of the policy but also the specific features of the assessment criteria. The research puts forward suggestions of improvements to be made to the ICAO policy and recommends ICAO to revise the LPRs at the earliest.

Acknowledgements

I would like to express my deep sense of gratitude to all my MA tutors for having assisted me in my distance learning process, especially my dissertation supervisor, Luke Harding, who has carefully guided me during the writing of this work, and Tineke Brunfaut, the Director of Studies of the programme, who has promptly assisted me whenever I needed her help. Very special thanks to the research participants, whose contribution was invaluable for the success of this study. Many thanks to the National Civil Aviation Agency – Brazil, the company I work for, which has kindly sponsored my study, and also to my coworkers for their continuous encouragement and support. I am also thankful to my family members and friends, who were very supportive and encouraging. I am especially grateful to my mother for having helped me transcribe the interviews and for having kindly reviewed this dissertation.

Table of Contents

Acronyms and abbreviations

AELTS	Aviation English Language Testing Service
ANAC	Civil Aviation National Agency
ATC	Air Traffic Controller
DECEA	Department of Airspace Control
DOC 9835	Document 9835 – Manual on the implementation of ICAO Language Proficiency Requirements
EASA	European Aviation Safety Agency
ELE	English Language Expert
ELF	English as a Lingua Franca
ELPAC	English Language Proficiency for Aeronautical Communication
EPLIS	The Brazilian Airspace Control System English Language Proficiency Exam
EUROCONTROL	The European Organization for the Safety of Air Navigation
ICAO	International Civil Aviation Organization
ICAO RSSTA	ICAO Rated Speech Samples Training Aid
ICAEA	International Civil Aviation English Association
I-HELPP	Homepage for the English Language Proficiency Programme
LPRs	Language Proficiency Requirements
LSP	Language for Specific Purposes
PRICESG	The Proficiency Requirements in Common English Study Group
RELTA	RMIT English Language Test for Aviation
RMIT	Royal Melbourne Institute of Technology
SARPS	Standards and Recommended Practices
SDEA	Santos Dumont English Assessment
SME	Subject Matter Expert
SOPs	Standard Operating Procedures
TLU	Target Language Use
UN	United Nations

1. Introduction

Pilots and air traffic controllers (ATCs)' insufficient language proficiency has been a contributing factor to several incidents and accidents that have happened in the history of civil aviation (ICAO, 2010). As pointed out by Kim and Elder (2014), "a single piece of unclarified information could have disastrous results in air traffic control" (p. 133). For this reason, the International Civil Aviation Organization (ICAO), a United Nations (UN) specialized agency, formulated the Assembly Resolution A32-16 in 1998, which urged the council to consider the matter of lack of proficiency in English by pilots and ATCs with a high level of priority and to complete the task of strengthening the regulations in order to obligate the Contracting States to take measures to ensure that pilots and ATCs would be proficient enough to conduct and understand radiotelephony communications in a safe way. The Proficiency Requirements in Common English Study Group (PRICESG) was established in 2000 in order to develop the ICAO Language Proficiency Requirements (LPRs). In March 2003, the council adopted the Standards and Recommended Practices (SARPs) concerning the LPRs.

I have been working with the implementation of the ICAO LPRs at the Brazilian civil aviation authority, the Civil Aviation National Agency (ANAC), since January 2008. Brazil, as an ICAO contracting State, needed to comply with the new language provisions. The government decided to develop and administer its own testing system to assess pilots and ATCs' language proficiency. Two different governmental organizations have been responsible for the implementation, development and administration of the Brazilian ICAO language proficiency tests: ANAC, responsible for the pilots' test (the *Santos Dumont English Assessment* – SDEA) and DECEA (Department of Airspace Control), responsible for the ATCs' test (the *Brazilian Airspace Control System English Language Proficiency Exam* - EPLIS). 12,165 pilots took the SDEA 24,203 times from December 2007 to August 2015. 15,504 final ratings were level 4 or above, and 8,699, level 3 or below. My work involves various activities such as regulation writing, State oversight, test developing, test administration, test conduction, rating, item writing and interlocutor/rater training.

I have chosen to research the ICAO LPRs to be the topic of my Masters dissertation because they are the basis of all the work that has been done, not only in Brazil but all around the world. Now, twelve years after the publication of the requirements, I consider it to be very important to know experienced professionals' opinions about the ICAO LPRs, what they consider to be the strengths and weaknesses of the policy, especially regarding the rating scale, learn from their experience and suggest ICAO ways to improve the policy, the starting point of the whole process. It is important to make it clear that the purpose of this study is not to criticise the ICAO LPRs for its flaws. In general, the policy is comprehensive and the guidelines are very helpful. However, although the publication of the ICAO LPRs has already been an important advance, there is still room for improvement. Hence, the ultimate objective of this study is to give ICAO feedback on the implementation of the ICAO LPRs from expert test developers and raters in order to assist the organization in revising the policy in order to make it even better. As Alderson, Clapham, and Wall (1995) argued, test content, administration, training and marking need to be a monitored ongoing process, so that they "can be modified and improved in the light of their performance and of research and feedback" (p. 218).

This dissertation consists of a qualitative study about the ICAO LPRs. Its aim is to investigate experienced raters and test developers' opinions about the ICAO LPRs in general, and, more specifically, the rating scale and the explanation of the descriptors. The participants were carefully selected. They are very experienced, interesting and unique subjects who have been involved with the best testing practices in the field. Although previous research has already indicated need for improvement of the LPRs and the rating scale, I do not know of any other studies whose aim was to collect recognized experts' opinions about the ICAO LPRs.

In the remainder of this introduction, I explain the ICAO LPR in more detail. Then, in chapter 2, I discuss what other authors have said about the ICAO LPRs. In this literature review, I first point out and discuss the main issues which have already been studied in relation to the LPRs in general. After that, I present the main issues that have been discussed related to the ICAO rating scale. In the end of the chapter, I introduce my research question. In chapter 3, I describe the overall research design, give information about the participants, and explain the type of data I collected, the data collection methods as well as the methods of data analysis. The results are presented in chapter 4 and discussed in chapter 5. The conclusion follows in chapter 6.

1.1 The ICAO language proficiency requirements

The ICAO SARPs require all aeroplane, airship, helicopter and powered-lift pilots, flight navigators, ATCs and aeronautical station operators working in international operations to "demonstrate the ability to speak and understand the language used for radiotelephony communications" (ICAO, 2010, Appendix A). This language can be the language normally used by the station on the ground or English. Initially, these aviation professionals should have their language proficiency level endorsed on their licenses as of 5 March 2008. However, since many Contracting States were not in compliance with these requirements by 2008, ICAO decided to extend the deadline to 5 March 2011, allowing for some flexibility over this deadline regarding the States that would not be able to comply with the LPRs by 2011.

According to the LPRs, tests should be designed to assess only speaking and listening and their purpose should be to assess plain language in an operational context. Phraseology, a standardized set of words and sentences used to ensure language used in radiotelephony communications is as clear and unambiguous as possible, should be tested separately from plain language. ICAO recommends that any errors involving misuse of phraseology or lack of technical knowledge should not interfere in the rating of the candidate's language proficiency. The target language use (TLU) domain should only be the English used in communications between pilots and ATCs. However, the Manual on the Implementation of ICAO Language Proficiency Requirements, from now on referred to as DOC 9835, explains that there are many different kinds of test tasks that can be used in order to elicit language. It talks about a narrow and a broad interpretation of work-related context tasks (both considered to be valid). Tasks which follow the narrow interpretation are restricted to replicating radiotelephony (including, of course, plain language), whereas tasks based on the broader interpretation elicit plain language on topics that vary from radiotelephony to aviation operations, including different kinds of situations, such as briefings, simulations, and role-plays.

The ICAO LPRs include the holistic descriptors (see Appendix A) and the ICAO rating scale (see Appendix B). The five holistic descriptors describe proficient speakers and the context for communications. Candidates who are awarded operational level 4 should have demonstrated compliance with the holistic descriptors. The ICAO rating scale consists of descriptors for six different categories (or skills) to be assessed:

pronunciation, structure, vocabulary, fluency, comprehension, and interactions. There are descriptors for six levels for each skill (from level 1, pre-elementary, to level 6, expert). The minimum level required for international operation is operational level 4. For a test taker to be awarded a level 4, he/she needs to have been awarded at least a level 4 in all skills, as the final level corresponds to the lowest of the six ratings, and not to the average of the ratings. ICAO recommended that a pilot or ATC who demonstrates language proficiency at level 4 should be formally evaluated at least once every three years. Professionals demonstrating proficiency at level 5 (extended level) should be evaluated at least once every six years. Candidates who are awarded level 6 (expert level) do not need to be assessed again. The rating should be done by at least two examiners, an English language expert (ELE) and a subject matter expert (SME). Because of the high stakes involved in this kind of testing, in case these two raters do not agree with the final level (between levels 3 and 4), the final result should be determined by a third rater. According to the policy, "native and very proficient non-native speakers with a dialect or accent intelligible to the international aeronautical community" do not need to go through formal evaluation (ICAO, 2010, Appendix A). They may be assessed by, for example, licensing authorities or flight examiners.

ICAO has taken some measures to help the Contracting States to implement the LPRs, including the publication of the DOC 9835 in 2004, and its second edition in 2010. This manual introduces the reader to the reasons why the LPRs were adopted, discusses basic concepts in language proficiency, language acquisition and language testing, explains the nature of radiotelephony communications, its general and specific features, details the language proficiency requirements (including an explanation of the rating scale descriptors) and provides guidance on the implementation of the LPRs. The organization has also published other important documents, such as the ICAO Circular 318, Language Testing Criteria for Global Harmonization, and Circular 323, Guidelines for Aviation English Training Programmes. Other actions taken by ICAO include the publishing of the ICAO Rated Speech Samples CD and the International Civil Aviation English Association (ICAEA)/ICAO rated speech samples training aid (RSSTA) (http://cfapp.icao.int/rssta/). ICAO also developed the Aviation English Language Testing Service (AELTS), which used to evaluate tests in order to check if they met the ICAO LPR. Although several tests had been submitted for this evaluation, only four tests were listed on ICAO's website (https://www4.icao.int/aelts/) as endorsed or conditionally endorsed tests. Only two tests are currently recognized by ICAO: the European Organization for the Safety of Air Navigation (EUROCONTROL) test, which is called *English Language Proficiency for Aeronautical Communication* (ELPAC) and the Royal Melbourne Institute of Technology (RMIT) test, the *RMIT English Language Test for Aviation* (RELTA). ICAO has recently stated that the AELTS has been suspended and that they are developing a new service, the ICAO Homepage for the English Language Proficiency Programme (I-HELPP).

Twelve years have passed since the publication of the ICAO LPRs. Although there have been some improvements in terms of manuals, projects, and services, the requirements themselves have not changed. There is controversy over the ICAO policy and the quality of its rating scale (Alderson, 2010, 2011; Douglas, 2004, 2014; Emery, 2014; Knoch, 2009, 2014; Farris, Trofimovich, Segalowitz & Gatbonton, 2008; Foy, 2012; Kim & Elder, 2009, 2014; Kim, 2013; Pfeiffer, 2009; Prado, 2015; Prinzo, 2009; Read & Knoch, 2009; Scaramucci, 2011). Alderson (2010) questioned: "are they (the ICAO scales) sufficiently explicit and relevant to guarantee that any test constructed on the basis of the ICAO scales will indeed be at the 'right' level or do the scales represent an uncertain and unstable foundation?" Douglas (2004) called for a revision of the ICAO LPRs in order to clarify areas of ambiguity and uncertainty. Kim and Elder (2009) noticed that there is strong resistance among Korean pilots towards the ICAO LPRs. Kim (2013) believes the non-native speakers' resistance towards the ICAO LPRs will only change after the ICAO policy and the construct underpinning it are revised. Knoch (2009) urged for more scale validation research by saying that her study "provides just one piece in the puzzle necessary to fully validate the ICAO rating scale" (p. 45). The need for more research is noticeable. Alderson (2009) named critiquing and revising the ICAO rating scale as a particularly important research area. As an effort to contribute to this discussion, the present study aims at finding out expert test developers and raters' opinions about the ICAO policy, in general, and the rating scale, more specifically.

2. Background

In this chapter, I present studies that are directly relevant to my research, introduce the research question and explain how my study adds to the body of knowledge. In the literature review section, I first talk about the main topics which have been discussed in relation to the ICAO LPRs in general. After that, I examine issues related to the ICAO rating scale, and then draw a few conclusions over what has been said.

2.1 Literature Review

2.1.1 Issues related to the ICAO LPRs in general

2.1.1.1 Characterizing pilot/ATC radiotelephony English as English for lingua franca (ELF)

Research on ELF have criticised the traditional understanding that native speakers own the language and that non-native speakers should speak the language according to the native-speakers' standards (Jenkins, 2000; Widdowson, 1994). Douglas (2014), Kim and Elder (2009), and McNamara (2012) all called for understanding the English used in radiotelephony communications between pilots and ATCs in the context of ELF. ICAO, in its DOC 9835, supports this view. In fact, McNamara (2012) identified one excerpt from the rating scale and three from the explanation of the descriptors which contain features of ELF context. However, ICAO contradicts itself. One contradiction lies in the fact that in spite of the organization stating that in the context of radiotelephony communications "it is no longer appropriate to use first-language or 'native' speakers as the model for pronunciation" (ICAO, 2010, section 2.5), the rating scale makes references to how much the candidate's pronunciation is influenced by the first language. According to the ELF theory, aiming for nativeness is unrealistic and unnecessary (Jenkins). Jenkins argued that the goal of communication should be intelligibility. Harding (2014) reinforced this idea by saying that the key in pronunciation assessment "is to focus on features crucial to intelligibility/comprehensibility, not nativeness" (slide 38).

Another part of the ICAO policy that contradicts the ELF view is that it puts the burden of effective communication on non-native speakers, in spite of its claim in the second edition of the DOC 9835 that "the burden of improved communication should not be seen as fallen solely on non-native speakers" (section 5.3). As previously discussed, native speakers do not need to be formally evaluated. As a result, the policy places the onus on non-native speakers of English to speak the language according to "the" standard (Kim & Elder, 2009; Read & Knoch, 2009; Scaramucci, 2011). As argued by Kim and Elder, both native and non-native speakers are accountable for communication problems, so all pilots and ATCs, regardless of their first language, should be trained to communicate effectively in English. Read and Knoch argued that the ICAO LPRs not only place the onus on non-native speakers to improve their proficiency, but "give native-speaking aviation personnel no incentive to develop their communicative competence in ELF terms" (p. 21.7).

2.1.1.2 Plain English versus phraseology

Scholars (Douglas, 2004; Moder & Halleck, 2009; Emery, 2014) are concerned about how to define the domain of English for radiotelephony communications. There is need to clarify the nature of what plain English is. Douglas (2004) claims it is "necessary to gather extensive and detailed information about the nature of aviation English, both the standardized phraseology and what is referred to as 'plain language', the relationship between them, and the conditions in which each is used" (p. 251).

Research has indicated that plain English tends to be favoured by speakers when they are dealing with abnormal or emergency situations, even when phraseology suffices (Kim & Elder, 2009; Morrow, Rodvold, & Lee. 1994; Howard, 2008). This underutilization of phraseology is a problem. By using plain English instead of phraseology, pilots and ATCs end up using more complex structure and vocabulary. As argued by Kim and Elder (2009), "plain English, in other words, is not very plain at all" (p. 23.14).

2.1.1.3 Lack of clarity in relation to the test construct

Research has shown that one of the biggest problems in the ICAO policy lies in the definition of what should be tested. Emery (2014) believes that "the ICAO guidance (…) is of little practical use in the definition of the test construct and the development of test specifications" (p. 206). The test construct vaguely defined by ICAO is ambiguous. For example, Douglas (2004) questioned the reason why face-to-face communications were included in the rating criteria despite the fact that pilots and ATCs only communicate on the radio. As very well put by one participant in Kim and Elder (2009)'s study, "if a test is to work properly, it should test actual radiotelephony communication and how well we cope with situations" (p. 138). I agree with that, as in real life, pilots and ATCs do not have to answer questions related to aviation, but to interact in radiotelephony communications. Nevertheless, it is important to point out that, although ICAO understands that the language to be assessed is just the language used by pilots and ATCs, it is likely that other kinds of communication might also have a significant impact on aviation safety, for instance, communications amongst the flight crew, between pilots and maintenance personnel, and so forth (Foy, 2012).

Some consequences of this lack of standardization in understanding the ICAO LPRs and unclarity in relation to the test construct may be pointed out: the policy has been implemented differently from country to country (Douglas, 2014); the validity, reliability, and meaningfulness of the ICAO language proficiency tests in the market are suspicious (Alderson, 2010); and some stakeholders' became reluctant towards the policy (Kim & Elder, 2014). Kim and Elder concluded that "what appears to be at the heart of the aviation experts' negative attitudes or resistance is the fact that the construct underpinning the test design fails to reflect critical features of radiotelephony communication including compliance with the radiotelephony conventions" (p. 145). In fact, the Korean test mentioned in their study lacks both situational and interactional authenticity (Kim and Elder, 2009). Indeed, just asking questions related to the relevant professional field does not make them authentic (Douglas, 2000). Nevertheless, it is not only the Korean test which presents problems (Alderson, 2010). I believe some of the problems come from deficiencies in the policy. As Douglas (2004) pointed out, for test providers to show that the interpretations that are being made based on the test results are justified, it is very important to have a clear picture of what is being assessed as well as a "clear, complete and unambiguous definition of the construct to be measured in relation to the purposes for which the measurement is being made" (p. 250). Unfortunately, the ICAO LPRs are not always clear.

The ICAO assessment criteria are not a good reflection of the TLU, as some irrelevant abilities were included, whereas some important ones were not taken into consideration (Douglas, 2014; Foy, 2012; Kim & Elder, 2009; Kim, 2013; Monteiro, 2012; Read & Knoch, 2009; Scaramucci, 2011). As Foy pointed out "we are testing (…) items of the English language that are not relevant to our everyday jobs" (slide 10). Knoch (2009) mentioned some aspects that seem to be irrelevant: idioms, comprehension of cultural and linguistic subtleties, and sensitivity to non-verbal cues. I personally agree that idioms and sensitivity to non-verbal cues should not have been included in the rating scale. However, based on Monteiro (2012)'s study results, which have been previously mentioned, I consider comprehension of cultural subtleties to be an important feature of the scale. On the other hand, Kim and Elder argued that ICAO oversimplified the test construct by not including the assessment of important communicative abilities. Kim claimed that "strategic competence for accommodation, and shared responsibility for lack of success of communication by participants should be incorporated into the radiotelephony communication construct and any tests which are designed to reflect this" (p. 107). Unfortunately, as pointed out by Douglas, interactional competence in the context of the ICAO LPRs is not assessed in any country. As he advocated, the assessment should test "linguistic awareness" and the ability to successfully manage a communication with a non-native speaker, especially the "abilities to accommodate their use of English in the context of intercultural communication" (p. 2). Similarly, Monteiro emphasized the relevance of raising pilots and ATCs' "awareness of the linguistic, discursive-interactional and intercultural factors" (p. 64) in order to improve radiotelephony communication safety. Foy also discussed the importance of training cultural sensitivity and highlighted that level 6 pilots tend to assume everybody understands them, so they do not employ strategies to make sure the messages are actually comprehended.

2.1.1.4 Not taking into consideration technical knowledge of operations in the language proficiency test

ICAO states that technical knowledge of operations should not be evaluated during the test. However, research (Davies, 2001; Ryan, 2007; Knoch, 2009) has indicated that it is difficult for subject matter experts to separate language ability from technical knowledge. Knoch pointed out that “it is possible that the testing of language and technical knowledge cannot or should not be separated” (p. 44). The results of her study also showed that, in relation to the criteria pilots use when rating other pilots, they take into consideration the following: “technical knowledge, experience and level of training”; “overall evaluation of level of speech”; “transition from standard phraseology to plain language”; “visual cues”; and “appropriacy of answers” (p. 37). The criterion most mentioned by the pilots when rating candidates was the first one, which was mentioned 50 times (23.26%). Douglas (2014) believes an ELF test needs to include background knowledge, as “the speakers must have something to talk about and knowhow to use that knowledge in specific situations” (p. 9). Emery (2014) summarizes this issue well by saying it is neither possible nor desirable to separate one from the other for three reasons:

> First, the test population is made up of licensed professionals known to be experts in the field. Second, the language policy itself requires that test content be field specific. Third, it is now generally accepted that the construct for LSP testing is one that allows for interaction between the test taker’s language ability and specific purpose language content knowledge. (p. 210)

2.1.2 Issues related to reliability

Garcia (2014), Knoch (2009), Pfeiffer (2009), and Scaramucci (2011) claimed that the ICAO rating scale may be interpreted differently by different raters. Knoch’s results showed that although ELE and SME raters rated most samples accordingly, there were some significant differences in their ratings. Accordingly, the results from Pfeiffer’s study, which investigated inter-rater reliability in a German test for ATCs, showed inconsistency among raters (low inter-rater reliability). Inter-rater reliability in rating comprehension was the lowest. The researcher observed “tendency towards the non-native speakers being less severe than the native speakers” (p. 40). Pfeiffer concluded that the main reason her reliability results were suboptimal was lack of common understanding of the ICAO rating scale descriptors. The results showed the raters were having difficulties in using the rating scale. The results from Garcia (2014)'s study showed low inter-rater reliability among the participants when rating pronunciation. The coefficients were lower among the participants who did not share the same testing context, which corroborates the idea that pronunciation assessment is very dependable on the raters' background.

Another reason for low reliability among raters might be related to the fact that some raters consider it to be difficult to separate one category from the other (Knoch, 2009, 2014; Pfeiffer, 2009). For instance, lack of vocabulary may be interpreted as low comprehension. Knoch (2009) expressed concern on how to deal with candidates who comprehend a question but cannot respond to it because of vocabulary problems. Another example would be the pronunciation criterion being weighed more heavily because raters do not understand something due to structure or vocabulary problems, and end up labelling the misunderstanding as a "pronunciation" issue (Knoch, 2014).

2.1.3 Issues related to the ICAO rating scale

There is little information available on the development and validation of the ICAO rating scale by the PRICESG (Kim & Elder, 2009; Knoch, 2009). As a matter of fact, not very much has been written about the use of the ICAO rating scale at all (Alderson, 2011; Pfeiffer, 2009). As pointed out by Emery (2014), "the scales have been in use in aviation language testing worldwide for more than 7 years, though there is no empirical evidence to support their validity for the high-stakes context in which they are used" (p. 207).

Furthermore, Knoch (2009) investigated three main issues: the stakeholders' opinions about the ICAO rating scale; if they agreed that operational level 4 was an appropriate level for international flying; and what criteria pilots use when rating other pilots. The results were inconclusive regarding the two last issues. Only around half of the participants agreed on level 4 being the adequate level for international operations, and 42% of the participants answered the rating scale was designed to be used by both ELEs and SMEs, 42% disagreed with it, and 15% were undecided. It seems aviation specialists have more difficulties in using the scale.

Speaking of the rating scale as a whole, the participants in Knoch (2009)'s study believe that the descriptors for level 6 are very similar to the descriptors for level 5, which makes it difficult to differentiate a level 5 from a level 6. Knoch suggested that the descriptors for level 6 need revision as they seem to be problematic.

Another recurrent theme in the literature about the rating scale was related to what skills are the most and the least important to be assessed. Some researchers believe that pronunciation is the, or one of the, most critical area(s) in pilots/ATCs radiotelephony communication safety (Brown, 1995; Kim & Elder, 2009; Knoch 2009, 2014; Ryan, 2007, Tiewtrakul & Fletcher, 2010). The pilots who participated in Knoch (2009)'s study focused more on pronunciation and fluency and less on structure. She believes that problems with pronunciation might seriously impact the comprehensibility of the message and they "often affect larger sections of a speech sample and therefore probably interfere with understanding more than, for example, individual wrongly chosen lexical items or slow speech" (Knoch, 2014, p. 85). Knoch also claimed that structure was not mentioned much, "one might conclude that correct structure is less important to the stakeholders in the TLU domain" (p. 42). Kim and Elder (2009) detected that pronunciation, comprehension and interactions are considered by stakeholders to be important criteria while structure, vocabulary and fluency are not. In the end, the results in Knoch's, and Kim and Elder's were similar in relation to the importance of testing pronunciation and the lack of importance of assessing structure, but they differed in relation to fluency.

Before I start discussing the issues regarding each of the categories, I recommend the reader to read the rating scale in Appendix B, the explanation of rating scale descriptors (Appendix C), or the summary I wrote about how the scale deals with each of the categories in Appendix D.

2.1.3.1 Pronunciation

A major criticism among the participants in Knoch (2009)'s study regards the reference made to first language influence. Knoch proposed the exclusion from the scale of any reference to the first language. Participants in her study also expressed concern about rating pronunciation because of the inclusion in the descriptors of adverbs of frequency, which were considered to be the only way of distinguishing between the levels. Moreover, a contentious issue among the participants was that rating pronunciation depends a lot on the background of the raters (i.e. the ones who work in multi-national environments usually have less problems in understanding candidates).

Pfeiffer (2009) believes the pronunciation rating scale descriptors are enough to rate candidates consistently, at least, in her German testing context. However, she said that "whether an accent or pronunciation mistake affects understanding seems to be a matter of personal perception and I think even in high quality rater-training measure this problem would indeed be very hard to tackle" (p. 43). A participant in Knoch (2009)'s study also criticised the term "understanding" as it is not clear who it refers to (understanding for whom?).

2.1.3.2 Structure

Participants in Knoch (2009) pointed out it cannot be assumed that unusual or unexpected situations elicit complex grammatical structures while routine situations elicit basic structures. Pfeiffer also acknowledged that "in the real aviation world, complex situations do not always generate complex language and hence the question arises whether the descriptor in question does not lack validity to a certain extent" (p. 16). A participant in her research also argued that "the degree of manipulation of basic and complex structures and the degree and density of grammatical error at level 4 to 6 is unclear" (p. 30). Nevertheless, their comments (which, as previously discussed, were few in relation to structure) were generally positive regarding this skill.

Pfeiffer (2009)'s perceptions about structure were less positive. She pinpointed several deficiencies in the structure descriptors. She believes that structure is the skill in which most discrepancies between standard phraseology and plain English appear, as the former requests simple language and the latter, more sophisticated structures. She claimed radiotelephony communications should be simple and that they do not require the elaborated level of speech described in levels 5 and 6. She pointed out this may lead to rater reliability problems, as the raters are given an unreliable instrument. Pfeiffer also argued that the definition "interferes with meaning" included in the scale is not very meaningful, as the understanding of what interferes with meaning or not is not uniform, not even among native speakers of the language. She identified several possible problems with rating structure which might negatively impact rater reliability. For example: what level to grade a candidate who controls basic grammatical structures well, but does not attempt complex structures; what to do when the candidate commits basic grammar mistakes, but attempts complex structures; how to deal with the rating scale lack of information about the amount of complex structures needed to grant levels 5 or 6 to candidates; when to downgrade a candidate due to the number of grammar mistakes. She feels that “a lack of common understanding how the ICAO level descriptors ought to be interpreted could lead to arbitrary judgments which might affect inter-rater reliability” (p.45).

Prado (2015) compiled a spoken corpus of the plain language used by pilots and ATCs in abnormal situations. She compared her results with the glossary of basic and complex structures published by ICAO in the second edition of DOC 9835 (Appendix E). She found out that in abnormal situations pilots and ATCs tend to repeat words (there are only a few relevant nouns in the corpus) and use simple language. The main verb tenses used are present continuous and simple present. Past simple, present perfect, simple future, and going to are less frequently used. This confirms Pfeiffer’s opinion that there is no need to use elaborated speech, as mentioned.

2.1.3.3 Vocabulary

One of the main issues in relation to the vocabulary descriptors relate to the fact that, although ICAO states in the explanation of the descriptors (ICAO, 2010) that “use of idioms is an obstacle to intelligibility and mutual understanding between non-expert users and should therefore be avoided by all users in this environment” (p. 4-11), the level 5 vocabulary descriptors say that “vocabulary is sometimes idiomatic” and in the level 6 descriptors that “vocabulary is idiomatic”. This needs urgent revision as pilots and ATCs should avoid using idiomatic expressions (Knoch, 2009; Pfeiffer, 2009). Although in other contexts, demonstrating knowledge of idiomatic expressions might show a more advanced proficiency in the language, it is not appropriate to assess idioms in a pilot/ATCs radiotelephony communication context. I strongly agree that idiomatic vocabulary should never have been included in the rating scale. Knoch recommends that "any references to idiomatic language should be deleted as this is not appropriate in the TLU domain" (p. 43).

Assessing sensitivity to register has also been spotted as a problematic feature of the rating scale (Knoch, 2009; Pfeiffer, 2009). Pfeiffer believes it is difficult to assess sensitivity to register since it is very dependable on the interlocutor's register. Knoch pointed out it is very influenced by the individuals' culture and suggested that the term "'register" should possibly be changed to "appropriacy" (p. 43).

Another issue in relation to the vocabulary descriptors refers to the meaning of "common, concrete and work-related topics" (Douglas, 2004; Knoch, 2009). A pilot who took part in Knoch's study pointed out that it is unclear whether it is one set of topics or three (p. 31). Douglas identified this unclear aspect of the rating scale as a potential threat to validity (p. 250).

2.1.3.4 Fluency

A recurring theme in the literature in relation to the fluency descriptors pertains to the assessment of use of discourse markers (Knoch, 2009; Pfeiffer, 2009; Prado, 2015). One participant in Knoch's study stated that discourse markers are not often used in the TLU and therefore should not have been included in the rating scale. This participant's opinion was upheld by Prado (2015)'s findings, that in most abnormal situations, pilots and ATCs usually either do not use any connectors or use simple connectors. She discovered that the connectors they use, when they use them, are: and, but, if, due to, because, and so. One of the pilots in the former study claimed to "have encountered speakers with high fluency who do not use discourse markers/connectors" (p. 30). Pfeiffer (2009) also acknowledged this problem and recommended that discourse markers should be thought of not in regard to their complexity, but their appropriacy.

Most participants in Knoch (2009)'s study were satisfied with the descriptors for fluency. However, there were some important comments. A participant argued that assessing variation of speech flow should not be taken into consideration as some speakers might be fluent without varying their speech flow for stylistic device. Both Knoch and Pfeiffer (2009) saw problems in this part of the descriptor. Another participant wondered if this should have been included in the pronunciation descriptors instead of fluency.

Knoch (2009) and Pfeiffer expressed concern about the definition of loss of fluency. Some participants in Knoch's study believe the words/phrases "distracting", "natural language flow", and "appropriate" included in the fluency descriptors are vague. One of them argued it is difficult to standardise the understanding of them as each person may have different interpretations of, for example, when fillers are being distractive. Luoma (2004) believes that raters should be more lenient in relation to the use of fillers when rating fluency, especially in spontaneous interaction, as using fillers is something natural. She argued that inattention might be the cause of mispronunciation and advocates that the focus should be on the effectiveness of the communication.

Another participant in Knoch (2009)'s study reported that the rating scale does not mention how well the candidates deal with the transition from phraseology to plain English, and that would be important to include. I believe this transition is what the scale writers meant by "transition from rehearsed or formulaic speech to spontaneous interaction", but they should have chosen the appropriate terminology. Another criticism was the absence of the definition of formulaic speech from the rating scale. Notwithstanding, I disagree with this criticism as this definition is included in the DOC 9835, a document which all raters should be familiar with. Knoch concluded that the fluency category "should only contain references to speed of speech, pausing, hesitations and possibly to fluency at the transition point between standard phraseology and plain language" (p. 42).

2.1.3.5 Comprehension

A recurring theme in Knoch (2009)'s study in relation to comprehension was that "comprehension could not be accurately measured in a scale designed to assess speaking performance" (p. 31), hence the difficulty to define how the comprehension scale would be operationalized. For example, one participant noted that some candidates might ask more questions than others because of their personality and this might make raters underrate them in comprehension. Moreover, Knoch (2009) and Douglas (2004) also expressed concern in relation to how intelligibility of a range of speech varieties (dialect and/or accent) in use internationally may be determined. I also find this problematic.

Although none of Pfeiffer (2009)'s colleagues mentioned having difficulties with the comprehension descriptors, the results of her study showed a different picture, as the inter-rater reliability coefficients for this category were the lowest. She believes that the reason for them to be lower in comprehension was that it is difficult to assess certain parts of the descriptors. Her feeling towards the comprehension descriptors is very negative. She criticises the PRICESG by saying that "the rating scale designers have not properly thought about the pertinency of the features to be included into the scale and hence a scale user could be seduced not to take the scale too seriously" (p. 56). She goes on saying that, in her opinion, the comprehension descriptors are "not very enlightening" and that, according to her judgment, "they are possibly the least well thought out in the entire rating scale" (p. 57).

2.1.3.6 Interactions

In relation to the descriptors for interactions, Emery (2014), Knoch (2009), and Pfeiffer saw the level 6 descriptor concerning sensitivity to non-verbal cues as problematic. Pfeiffer stated that "I naturally cannot rate non-verbal cues with only a tape recording at hand" (p. 18). She expressed concern that there might be differences in rating interactions live and from test recordings. A number of participants in Knoch (2009)'s study also believe that sensitivity to non-verbal cues should be excluded from the rating scale as they are irrelevant for radiotelephony communications. Emery argued that it would be problematic to operationalise the assessment of this ability.

Pfeiffer (2009) brought up another issue related to the level 6 interactions descriptors. She claimed that it is not clear what the scale means by saying the candidate "interacts with ease", as the raters in her research showed different interpretations of what it is to interact with ease.

Knoch (2009)'s study indicated that the descriptors for level 5 and level 6 might have been swapped as the level 5 descriptors seem to describe more advanced abilities than the descriptors for level 6. Additionally, there was criticism regarding the fact that some features do not appear across all levels. One of the participants gave an example of a feature he thought should be included in all levels: the ability to deal with apparent misunderstanding by checking, confirming or clarifying (from level 4).

To finalise, Douglas (2014) suggested that the interactions category should be revised and should gain prominence. He believes more features of ELF should be included for the assessment to reflect the real ELF context.

2.1.4 So what is next?

Based on this concise review, we may conclude that the ICAO policy and the rating scale need to be revised. Prinzo (2009) suggested that more research is necessary to "pave the way for further revision of the ICAO rating scale descriptors" (p. 9). Knoch (2009)'s study results showed that the participants were generally satisfied with the scale categories and descriptors. However, despite this general acceptance, the open-ended responses pointed out that revisions might be necessary. Farris et al (2008) and Knoch (2009) recommended there should be more research to validate the ICAO rating scale. Pfeiffer (2009) concluded that "the ICAO rating scale obviously does not offer enough help to come to reliable judgments and needs amendments" (p. 60).

Prinzo (2009) argued that raters would be more reliable if they had quantitative information in the scale to rate candidates. Farris et al (2008) also believe that it would be useful if ICAO developed more objective measures of language proficiency. They recommended improvements to be made to the ICAO rating scale in terms of objectivity in order to increase the scale reliability. In their opinion, the scale should be more objective and related to/associated with the communicative feature of the pilots/ATCs' communications. However, Pfeiffer (2009) disagreed that the insertion of quantifiable metrics would be a solution to the problems with rater reliability. She claimed that, for instance, by saying a candidate's pronunciation only interferes with the ease of understanding a certain percentage of the time does not mean much, as the raters are not going to count the number of words the candidate mispronounced to know exactly how many times his/her pronunciation interfered with ease of understanding. However, she considered inserting the percentages into the scale to be better than not taking any action. In addition, Pfeiffer acknowledged that it is unlikely that the rating scale will ever be interpreted in a uniform manner worldwide. She suggested that each test provider should establish a comprehensive agreement on how to interpret the descriptors. Some participants in Knoch (2009)'s study suggested that there should be an international rater network as an effort to try to standardize the interpretation of the descriptors amongst raters around the globe.

Furthermore, participants in Knoch (2009)'s research suggested that ICAO should audit tests. Indeed, some authors have expressed deep concerns whether the tests that have been used to assess pilots and ATCs around the world follow good standards (Alderson, 2009, 2010, 2011; Huhta, 2009; Read & Knoch, 2009). These concerns should be taken seriously as this is high-stakes testing. As argued by Alderson (2011), "if the language proficiency tests used to license aviation personnel are unreliable or lacking in validity, there are potentially dangerous consequences". (p. 389). However, although ICAO, as mentioned, has made the AELTS available for some time for the test providers who wanted to go through the endorsement process, the approval was not obligatory. ICAO claims that the responsibility for approving language tests falls on each State's civil aviation authority (from: http://www.icao.int/safety/airnavigation/pages/peltrgfaq.aspx#anchor17).

As seen in this literature review, there are many polemical issues involving the ICAO policy. I have summarized the recurrent topics and the ones I consider to be the most significant. Although the ICAO LPRs may still be in its infancy (Emery, 2014), research has shown they already need to go under the knife.

2.2 Research question

Research has been done to investigate important aspects of the ICAO LPRs, such as the quality of ICAO language proficiency tests, inter-rater reliability, stakeholders' opinions about the rating scale, what stakeholders consider important to be assessed, the nature of miscommunication in abnormal and emergency situations, among others. However, I have not heard of any study that researched either recognized test developers' opinions about the policy or experienced ELE raters' perceptions of the rating scale. With the present study, I intend to answer the following research question:

- What do recognized ICAO test developers and experienced raters perceive as the strengths and weaknesses of the ICAO language proficiency requirements?

This research question refers not only to general topics related to the ICAO policy, such as the ones discussed in the first section of the literature review, but also to the rating scale and the explanation of the descriptors, such as the topics discussed in the second part of the literature review. With this question I intend to learn what aspects of the ICAO policy have been working well, as well as the difficulties test developers and raters have been experiencing. The extension of this is to find out what parts of the policy should be maintained and what should be changed. With that, I would like to add to the body of knowledge a list of suggestions and recommendations for improvements to be made to the ICAO policy in general and to the rating scale.

3. Methodology

3.1 Overall research design

This study consists of a primary qualitative research project, which involved the participation of six test developers and raters. One method (interviews) and two different sources (five ELEs and one SME) were used. Semi-structured interviews were carried out in order to investigate the participants' subjective and detailed opinions about the ICAO LPRs. An interview guide (see Appendix F) comprising 15 questions about the ICAO policy (part 1 of the interview) and 16 questions about the rating scale and the explanations of the descriptors (part 2 of the interview), plus two final questions, was developed to help the collection of rich and in-depth data. Part 1 of the interview was mostly comprised of questions asking the participants if they agreed or not with the main features of the policy and why. The second part of the interview focused on the strengths and weaknesses of each of the rating scale descriptors and the explanation of the descriptors. The language used in the interviews was English. They were audio-recorded and transcribed for data analysis. The data were analysed following the methods described in section 3.3. See Appendix G for an interview sample.

3.2 Participants

The six experienced test developers and raters who participated in this survey did so on a voluntary basis. They were selected according to a purposive sampling technique (as explained in Dörnyei, 2007). They have all engaged with the ICAO LPRs in a lot of detail. Five of them have been involved with both test development and rating. Five are ELEs and one is a SME. Three of them hold a Masters Degree in Linguistics or Applied Linguistics and two of them are currently doing their PhD. They can be considered unique participants since four are or have been involved with the development of tests or rating in tests which have been endorsed by ICAO (three out of four tests that have already been endorsed by ICAO). Moreover, one participant was a member of the PRICESG. Two participants have coordinated the ICAO rated speech samples project, and other two have participated in it as raters. As a matter of fact, one participant is an ICAO test evaluator. Furthermore, four of the six participants are board members of the International Civil Aviation English Association (ICAEA). Four participants have been engaged in organising and lecturing at important international conferences in the field, and two have published related articles in scientific journals. Two participants have also been involved with regulation writing as they work for their State's civil aviation authority. Also, half of the participants have experience with training pilots and/or ATCs. One of them is a reputable author of training material. They come from six different countries, three in Europe, two in South America and one in Oceania. Three participants were male and three, female. Their ages were 39, 40, 46, 53, 54 and 69 (average age of 50 years old). They have been working with the ICAO LPRs for 8, 9, 10, 11, 11 and 15 years (average of 10.7 years). I have decided not to include a table with the background information about each participant because of confidentiality issues.

3.3 Type of data and data collection methods

Rich and complex qualitative data were collected through one-on-one interviews between June and September 2015. As participants were from six different countries, the interviews took place through Skype. Half of the interviews were conducted in voice only, and half with video, depending on the participants' choice. The interviews varied in duration, taking from 55 minutes to 2 hours 29 minutes (see Table 2 for duration of interviews). The interviews were audio recorded and the data were transcribed verbatim. The purpose of the interview was to gather intense, full and saturated opinions from the participants (as described in Polkinghorne, 2005).

Table 1

Duration of interviews

Participant	Duration of the interview
A	1h20
B	2h29
C	1h10
D	1h09
E	0h55

F	0h41 *

Note: * The interview with participant 6 was atypical. He had to leave and only answered the questions from the first part of the interview and the first question of the second part (we tried to schedule a second meeting to finish it, but, unfortunately, the participant did not have availability).

3.4 Methods of data analysis

The interviews were transcribed and the data were analysed in three phases. First, the answers to the questions from the first part of the interview were analysed in regard to how many participants agreed/disagreed with each feature of the policy and what their arguments were. Then, the answers to the questions about the rating scale were analysed in relation to what the participants considered to be the descriptors main strengths and weaknesses. Finally, a thematic analysis was undertaken in order to determine what other prevalent and relevant themes in the interviews were. The thematic analysis followed the guidelines on using this kind of approach which were described by Braun and Clarke (2006). I “generated initial codes” and collated the relevant excerpts below each code. After that, I “searched for themes” and “reviewed them”. Next, I “defined and named” the prevalent themes. Finally, I selected the best quotations to illustrate participants’ arguments. The quotations I chose to include in this dissertation are the ones I think represent the essence of the point the participants were trying to make.

I have decided to use thematic analysis to analyse the data because it is a very useful, flexible, and straightforward method in qualitative research, usually easy to be conducted by novice researchers (Braun & Clarke, 2006) and also because I consider it to be appropriate to the purpose of answering my research question. When analysing the data, as suggested by Cohen, Morrison and Manion (2007), I took special care in order not to forget the synergy of the whole when separating the data into fragmented elements. As they pointed out, “the great tension in data analysis is between maintaining a sense of the holism of the interview and the tendency for analysis to atomize and fragment the data” (p. 368).

4. Results

The main findings from each phase of the analysis are introduced in this chapter, which is divided into three sections, as shown in Table 2.

Table 2

Chapter 4 sections

Section	Subtitles
4.1	Participants' opinions about the main features of the ICAO policy
4.2	Participants' opinions about the strengths and weaknesses in the assessment criteria
4.3	Recurring/relevant themes related to the ICAO LPRs

4.1 Participants' opinions about the main features of the ICAO policy

Table 3 shows the main features of the ICAO policy that were discussed in the interviews.

Table 3

Main ICAO policy features

Number	Description of the policy feature
1	The target language use domain is the English used in communications between pilots and ATCs.
2	The tests should be designed to assess speaking and listening
3	The purpose of the test is to assess plain language proficiency in an operational aviation context.
4	Phraseology should be tested separately from plain language.
5	Responses containing elements of ICAO phraseology should not be rated with regard to their procedural appropriateness or technical correctness.
6	Technical knowledge of operations should not be evaluated.
7	Operational level 4 is enough for safe operations.
8	Those demonstrating language proficiency at the operational level 4 should be evaluated at least once every three years.
9	Those demonstrating language proficiency at the extended level 5 should be evaluated at least once every six years.
10	Expert Level 6 candidates do not need to be tested again.
11	A candidate who is tentatively considered to be a level 6 speaker of the language may be evaluated through informal assessment (for example, by a flight examiner or licensing authority).
12	The six categories that need to be assessed are: pronunciation, structure, vocabulary, fluency, comprehension and interactions.

13	A candidate's final level should be the lowest level in any of the categories.

I have numbered each feature of the policy in Table 3 to make it easier to refer to them in Table 4. Table 4 provides a summary of the participants' responses to the question if they agreed with each feature of the policy or not.

Table 4

Participants' responses to the main features of the policy

ICAO policy	**A**	**B**	**C**	**D**	**E**	**F**
1	Agrees	Disagrees	Agrees	Disagrees	There should be a test whenever English is required	There should be a test for everything that is RT
2	Mostly agrees	Disagrees	Agrees	Mostly agrees	Disagrees	Agrees for the time being
3	Agrees	Disagrees	Disagrees	Agrees	Agrees	Agrees
4	Agrees	Unsure	Disagrees	Agrees	Agrees	Agrees
5	Agrees	Agrees in part	Disagrees	Agrees	Disagrees	Agrees
6	Agrees	Agrees	Disagrees	Agrees	Agrees	Agrees
7	Unsure	Disagrees	Unsure	Disagrees	Agrees	Agrees
8	Agrees	Disagrees	Unsure	Disagrees	Disagrees	Agrees
9	Agrees	Disagrees	Unsure	Disagrees	Disagrees	Disagrees
10	Disagrees	Disagrees	Disagrees	Disagrees	Disagrees	Disagrees
11	Disagrees	Disagrees	Disagrees	Disagrees	Disagrees	Disagrees
12	Agrees	Mostly agrees	Mostly agrees	Agrees	Agrees	Mostly agrees
13	Agrees	Agrees	Agrees	Agrees	Agrees	Agrees

In relation to the ICAO policy feature number 1, the participants usually agreed that testing pilots and ATCs' ability to communicate in English on the radio is the most critical area in relation to safety. However, many participants acknowledged the importance of testing other abilities as well. Participant A said "controllers (…) sit across the room and talk to a colleague. These days (…) you have different nationalities and more so with the pilots, they talk in the cockpit. It is very important for pilots to have a high level of English. They can communicate to flight engineers, to cabin crew, to cockpit crew, but we are not testing it”. Participant B also argued that “interactions between pilots inside the cockpit are vital. They are very critical and they can really cause problems in the operations”. Participant F mentioned “that is something that needs to be taken into consideration, especially if you keep in mind that, if you take an airline like Emirates, on the Airbus 380, Emirates may have 27 cabin crew, and that means 27 languages”, and added that “for everything that is radiotelephony, that is non-visual, I think it’s good to have a specific test, yes”. Participant D thinks it would also be important to train and test mechanics “because of the high stakes which are involved in their activity. However, half of the participants believe expanding the target language would be very complicated. As participant C argued, “they (the other target use domains) are all so varied and so different, that it would be difficult to capture the essence of all of those multiple uses of language and all of the variation in one set of criteria for assessment”.

Regarding policy number 2, the participants were not very confident that the tests should be designed to assess only speaking and listening. The possible need for testing writing because of the advance of datalink was mentioned by most of them. Participants B and E also considered reading to be important. The former advocated: “Reading is an important skill for pilots because they read checklists, they read manuals, they read SOPs, they read lots of documents during the flight. (…) They use these documents before the flight, during the flight, and after the flight. So if they are not skilled in reading, you could have a problem”. Participant E included that “writing would help a pilot’s proficiency overall”. On the other hand, two participants expressed concern in relation to practicality issues involved in assessing reading and writing.

Two participants agreed with the third feature of the policy, which says that the purpose of the test is to assess plain language proficiency in an operational context, whereas two agreed in part, and two disagreed. As argued by participant D, “we have experienced some cases with pilots and controllers with excellent level of general English, and when they were tested, they were not considered to be operational at all. So having a good level of general English doesn’t mean that they will be operational. Being operational, managing the language in an operational context, is completely different. (…) They should be trained in dealing with aviation English, plain English in the aviation context. It is essential". Although participant F agreed with this part of the policy, he expressed concern in relation to the clarity of the terms. He said "I agree that the purpose of the test should be to assess plain language proficiency in an operational aviation context, but the question is: what is an operational aviation context? That is where there is a lot of discrepancy. (…) The operational context of aviation is not so easily defined. (…) It's very easy to say the operational requirements of aviation, but what exactly that means is a very different thing". Although participant E also agreed with the purpose of the test which was stipulated by ICAO, she advocated that candidates should only be able to qualify for the ICAO test after being approved in a phraseology test. Participants B and C claimed that the purpose of the test should be to assess plain English as well as phraseology. As thoroughly discussed by the latter:

> No, I don't agree with that. I believe that the intention of ICAO is to assess language proficiency in the context of radio communications and that target language use domain is made up of two really important components, first being standard radiotelephony phraseology and the second being plain English where phraseology doesn't suffice. So I think that ICAO missed an opportunity to combine those two elements, the target language use domain. And I think had they done that, had they encouraged test developers to work on both phraseology and plain language together, then, on one hand, we would have stronger tests because naturally you would want to engage the ability of the candidate to improve and use, demonstrate their ability to use both standard radiotelephony phraseology and plain English, so the quality of test would be better and would be linked more to air/ground communications, and I think, secondly, it would have helped ICAO to address the issue of proficient uses of English insofar as I quite strongly believe that a lot of communication problems are to do with poor use of phraseology amongst native speaking crew and possibly air traffic control too. (…) So I think it should include both phraseology and plain English. To separate the two is to artificially divide a single construct which is safe pilot/controller communications into constituent parts which don't necessarily want to be divided.

The previous topic is closely related to the next one, which is about plain English being tested separately from phraseology. Although the majority of participants agreed that it is better to separate them into two different tests, participant B was not sure about it and participant C disagreed with it (as it can be seen in the quotation from the previous paragraph). Participant B claimed it would be difficult to assess them both in a single test, whereas participant F agreed phraseology needs to be tested separately from plain English because "they are very different codes". In any case, all participants agreed that English phraseology needs to be formally tested. Two of them showed concern for it not being tested among pilots in their countries, and another participant expressed concern for it only being tested at the beginning of the career. Moreover, they all agreed that phraseology should be included in the language proficiency test, especially, as argued by two participants, the transition from phraseology to plain English.

With respect to not rating responses containing elements of ICAO phraseology with regard to their procedural appropriateness or technical correctness, three participants agreed with this part of the policy, whereas one agreed in part and two disagreed. Participant E disagreed with it because, according to her, the two integrate a lot in real life, which should be reflected by the test. Participant C also disagreed and argued that "the ability to communicate effectively in a radiotelephony context is dependent on A) accurate use and appropriate use of standard radiotelephony phraseology and B) where that phraseology doesn't suffice, good command of concise, brief, and clear plain English. So I think being a proficient user of the radio requires skills on both sides and I think tests should be measuring both of those things in tandem together because they are part and parcel of the same construct". Most participants argued that, in case weaknesses regarding the use of phraseology are detected, action should be taken. Participant E argued that "it needs to be noted and immediate follow-up given somehow". In participant A's testing context, this action is to bring it to the attention of senior managers, and in participant D's context, raters discuss the problems that were noticed and make recommendations to the candidate.

Five participants agreed that lack of technical knowledge of operations should not be taken into consideration by the examiners when rating. Participant A claimed that as it is not a professional test, we cannot evaluate the candidates' abilities to fly or control an aircraft. Participant D agreed these two kinds of assessment should not be mixed. However, although participants B and D agreed with it, they both acknowledged it is difficult to dissociate technical knowledge from linguistic knowledge, especially for SME raters, "because the context is so specific that this technical knowledge is almost intrinsic to the language" (participant B). Participant D talked about the importance of making it clear during training courses that raters should not be influenced by the candidates' lack of technical knowledge. Participant A, again, recommended that if the SME rater identifies errors, he would have to report it. Participant D said they may, in the score report, require the candidate who made noticeable mistakes to undergo recurrent training. Participant F shared the information that the reporting forms used by the raters in his testing context contain a box for the SME to make comments on these issues. On the other hand, participant C, who was the only one that disagreed with this feature of the policy, argued the following:

> I think it's impossible not to evaluate technical knowledge during a test, for example, if you have in a well-developed task which simulates radiotelephony communications, if you give a pilot or a controller a scenario which engages plain language use in that context, you cannot separate procedural or operational knowledge from that language performance. For example, if a pilot is talking about hydraulic loss, engine problems or weather issues, the way that he or she chooses vocabulary, chooses how to wrap up their meaning into plain English communications, will be dictated by their knowledge of operational procedure. (…) It is very specific and it is very linked to the operational context in which the traffic is operated. So I think it's impossible and undesirable to try and separate operational knowledge from language knowledge.

Concerning the operational level 4 being enough for safe operations, the participants' opinions varied. Two of them agreed that level 4 is enough, two disagreed and two were uncertain about it. Participant D believes the descriptors for level 4 allow for too many weaknesses. Participant B thinks level 4 is sufficient for routine situations, but not for emergency situations, because "when you are testing the candidate, although they feel nervous, they are not really executing their tasks. In non-routine, emergency situations they will have a lot more to deal with, so they will become nervous, they will get stressed, they will have to deal with flying the airplane, and doing so many other things, coordinating with the crew to deal with an emergency, or an abnormal situation that language will be their last concern. (…) Unintentionally, his language abilities will decrease".

Half of the participants disagreed that three years should be the maximum interval of assessment for candidates who get level 4. One participant was not sure, and two participants think it is an appropriate interval. Participant B believes a level 4 candidate can lose proficiency in English quickly if he/she does not keep in contact with the language. She argues that candidates at this level should be retested after two years, at the most. Accordingly, participant D also believes that level 4 candidates should be retested after two years. Both participants B and E have noticed level 4s deteriorating to level 3 after three years. Three participants believe that reducing this interval to two years would encourage the candidates to keep studying in order to maintain their language proficiency. Although participant F believes three years to be "ok", he mentioned he finds "slightly problematic how they came up with three years". Furthermore, two participants mentioned that the European Aviation Safety Agency

(EASA) is in the process of changing the interval of assessment for level 4 to four years in order to fit the language assessment into their testing cycle. They both criticised this decision.

Four participants disagreed with the maximum interval of assessment of six years for level 5 candidates, one expressed uncertainty, and one agreed with it. Participant B believes the interval of six years for reassessment of level 5 candidates is too long, as, although some pilots in her testing scenario kept the same level and very few performed better than level 5 on the reassessment, she has seen many going down to level 4, and some even going down to level 3. She claimed that this interval should be reduced to a maximum of five years. Participants D and E suggested level 5 candidates should be retested after four years. As argued by participant E, "level 5s can still make errors in complex structure. They are not perfect by any means, so maybe four years would be ideal. (…) But anyway it would make pilots and ATCs aware that you can't just let it go. It is an ongoing process". Moreover, participant F argued six years "is more or less a random number" and advocates that everybody should be tested again after three years "because I think even if you have very good language abilities, three years is enough time for you to lose a very large chunk of your language if you don't use it every day. The ATCs are always using more or less the same parts of their language register, so they lose it fairly quickly".

None of the participants agreed that level 6 candidates do not need to be tested again. Participant A thinks this is "the worst thing ICAO did". All participants claimed that everybody needs to be retested and gave various examples of how candidates can lose language proficiency over time.

Again, none of the participants agreed that candidates who are tentatively considered to be level 6 speakers of the language could be evaluated through informal assessment. They all believe it has to be a formal test. Many participants expressed concern that language proficiency needs to be assessed by a trained rater. Participant E argued that "the person testing might not be that good themselves", whereas participant B questioned how a flight examiner can judge pilots' language proficiency and know when to direct them to a formal assessment. Moreover, participant E and F argued that the assessment criteria need to be clear. The latter advocated that: "the test must be standardized. If you do it as an informal test, it cannot possibly be standardized". Participant C also develops a very good argument:

> Language tests are developed with a specific purpose in mind and that in our context is to make valid influences about the ability of a pilot or a controller to communicate effectively. So, I think that an informal testing context will not engage the abilities that we are looking for in safe radiotelephony communications (…). If, for example, you are in a simulator, you do a series of simulator exercises and then you have a debrief with your simulator instructor in order to be signed off and your license revalidated, the discussion you have with that instructor is not going to engage the range of listening comprehension abilities or the ability to switch between standard phraseology and plain English, so, no, I strongly disagree that informal testing of high level users of English is an acceptable means of determining the ability to communicate safely on the radio.

The participants mostly agreed that the six areas that should be tested are the ones that were included in the rating scale. However, participant B believes that the descriptors "are not enough for the complexity of this communication", and argued that there are abilities that need to be included, for example, "features of cultural competence". Similarly, participant F believes the descriptors are not enough as they omit the assessment of skills such as "avoiding ambiguity, avoiding idiomatic expressions, accommodating a weaker speaker". Participant C criticised the fact that comprehension is only one of the six criteria in the rating scale because "it makes us think that comprehension is less than 20% of the overall ability to communicate on the radio. It is not, it is 50%, at least". He advocated that comprehension should have its own scale because:

> Comprehension sitting alongside components of spoken language proficiency firstly diminishes the importance of listening comprehension. If we consider listening/speaking to be skills which are equal and they interact and relate very closely together the way that comprehension is perceived in the rating scale is perceived as one of six things that students should be able to do, when it is not, it is one of two things that (…) pilots and controllers should be able to do, one being speaking and the second being comprehension. So I think it misleads us to think that comprehension is a very thin slice of the ability to speak, and it is not. It is an extremely important, if not more important, part of the overall proficiency construct in this case.

As a matter of fact, ICAO added to the second edition of the DOC 9835 the following sentence: "while comprehension is only one out of six skills in the Rating Scale, it represents half of the linguistic workload in spoken communications" (ICAO, 2010, 4-13).

All participants agreed that the candidates' final level should be the lowest level in any of the categories. The arguments usually reinforced that this is a very high stakes testing context and weaknesses in one area could lead to serious misunderstandings. Participant C believes it not only improves safety, but "it helps to improve test reliability, in that two examiners are much more likely to agree on the overall operational proficiency of a candidate when you take into account the lowest of any of the scores in each of the criteria".

4.2 Participants' opinions about the strengths and weaknesses in the assessment criteria

All participants talked about positive and negative aspects of the rating scale. Participant A is mostly satisfied with the descriptors and pointed only to a few adjustments. Four participants are partially satisfied with them and indicated the need for some changes. Participant C believes the scale is "very poorly thought out" and "really poorly defined". This section presents the recurring strengths and weaknesses regarding pronunciation, structure, vocabulary, fluency, comprehension and interactions. Appendix H lists less recurring but relevant strengths and weaknesses brought up by participants regarding all categories.

4.2.1 Pronunciation

Table 5 shows the recurring strengths and weaknesses discussed by the participants regarding pronunciation.

Table 5

Main strength and weaknesses related to the descriptors for pronunciation

Strength	Weaknesses
Focus on comprehensibility	Focus on how much pronunciation is influenced by the first language
	Use of adverbs of frequency as a measure to assess how much a candidate's pronunciation interferes with the ease of understanding

As discussed earlier in this dissertation, one of the main criticisms regarding the pronunciation descriptors lies in the fact that they include the assessment of how much a candidate's pronunciation is influenced by the first or regional variation. They claimed this does not matter, what matters is comprehensibility.

On the other hand, most participants argued that a very positive aspect of the descriptors is the fact that they talk about how much the candidates' pronunciation interferes with understanding. However, all participants criticised the fact that the difference between the levels is based on frequency. Adverbs of frequency were considered to be a very subjective way of assessing interference with understanding, especially because the descriptors for pronunciation are very similar across all levels. As explained by participant B, "depending on the level, they are so similar that we can't make a clear distinction". Most participants disapproved of this subjectivity, like participant F, who said: "These are subjective value judgments. What I perceive as being 'rarely' may not be the same thing you think is rarely". Participant E also pointed out that "it doesn't really account for when there is one error that is really bad (…) occasionally there is one really, really bad error that makes you 'oh my goodness, that needs addressing', (…) it's based on frequency, it doesn't cater for where". Participant D suggested that "we should find another linguistic way of expressing what is required for each of the levels. Adverbs of frequency can be misinterpreted with other similar adverbs. This is where different interpretations appear". As an effort to help assessors rate candidates more objectively, in participant A's testing context they have included percentages for each of the levels as a guideline.

4.2.2 Structure

Table 6 shows the recurring strengths and weaknesses discussed by the participants regarding structure.

Table 6

Main strength and weaknesses related to the descriptors for structure

Strength	Weaknesses

Focus on interference with meaning	Difficulty to differentiate candidates' level based on control of basic and complex structures
	Difficulty to work with the glossary of basic and complex structures

The candidates usually commented positively about the assessment criteria for structure. A number of participants mentioned that a positive aspect of the descriptors for structure is the focus on global and local errors, in other words, on the interference with meaning. Participant C stated that he likes "the way the focus is on global and local error", and participant B believes "this part of the descriptors is very important". However, it was argued that there is need for clarification of what interferes with meaning and what does not.

One of the main weaknesses regarding structure is that some participants experience difficulty to fit some candidates in the criteria. Participant C stated that:

> There is a very delicate and poorly understood interplay between basic and complex structures, for example, at level 5, it says basic structures need to be consistently well controlled with complex structures with error. It's very rare to come across candidates who do that. It's much more frequent to come across candidates who consistently make few errors with their basic structures and attempt complex structures and sometimes they get that right as well, but it's still got error in it.

Similarly participant E pointed out that "you actually often get pilots who are trying complex structures all the time, but the basic ones aren't well controlled, which pushes them down to 4".

Another recurring negative comment was about the glossary of basic and complex structures. Participant C argued that "the list of basic and complex structures (…) is not rooted in any sense of research either in the target language use domain or in the wider field of applied linguistics language teaching, and I think it's really poorly thought out". He illustrated that it includes language features that are more related to vocabulary than to grammar (i.e. gradable and ungradable adjectives). Participant B mentioned that some of the structures that she used to consider complex were included as basic, for example, "passive voice, present perfect, present perfect continuous, some modals like 'ought to', some relative pronouns, the position of direct and indirect objects". The reason for this might be, as participant C pointed, the fact that "it is actually extremely difficult to categorise structures as basic or complex particularly when you're dealing with international uses of English where for some speakers a particular structure in English may be highly complex because there is no equivalent in their first language". Furthermore, participant B also pointed out that some discourse markers (which appear in the rating scale as a feature of fluency, not structure) should never be used in radiotelephony, for instance, "on top of that", "in short", "nevertheless", "mind you" and "by and large". She added that "if they (pilots and ATCs) used some of those connectors, it could cause misunderstandings".

4.2.3 Vocabulary

Table 7 shows the main recurring strength and weaknesses discussed by the participants regarding vocabulary.

Table 7

Main strength and weaknesses related to the descriptors for vocabulary

Strength	**Weaknesses**
Reference to the ability to paraphrase	Reference to idiomatic vocabulary
	Reference to sensitivity to register

The main strength identified by most participants is the reference to the ability to paraphrase, which is very important in this context. They claimed that the two outstanding weaknesses in relation to the vocabulary descriptors are the references to idiomatic vocabulary and sensitivity to register. They argued that idioms should be avoided in pilot/ATCs radiotelephony communications, and that the register in this kind of communication does not vary much.

4.2.4 Fluency

Table 8 shows the only recurring weakness discussed by the participants regarding fluency.

Table 8

Weakness related to the explanation about the descriptors for fluency

Weakness
Confusion regarding the understanding of the explanation about the ICAO recommended rate of 100 words per minute

Four of the five participants who answered the questions related to interactions mentioned problems regarding the explanations for levels 4 and 5, which describe that the speaker has the possibility of speaking a little faster (level 4) or significantly higher (level 5) than the ICAO recommended rate of 100 words per minute. Participant B argued that this feature of the explanations is “questionable”, but that it might help in case of doubt. Participant E believes that this is not helpful, whereas participant D does not feel comfortable with this feature because she has experienced cases where the candidates did not speak as fast, but the performances were satisfactory. On the contrary, participant C believes this rate is “unnaturally slow” and that pilots and ATCs “never do it”.

Additionally, another recurring theme regarding fluency was related to the part of levels 5 and 6 descriptors which talks about variation of speech flow as a stylistic device. However, it is not possible to distinguish if this feature is a strength or a weakness as two participants believe it to be important, whereas one participant thinks it is not. Participant B considered these descriptors important because “there are times that you really need to emphasize your information. This is a good point, but not as a stylistic device, it is not just for style, it’s just for urgency or emergency, you need to emphasize for a reason and in this context we have some moments that this is really necessary”. Participant C believes it is important too, but thinks it should have been included in the descriptors for pronunciation. On the other hand, participant E reckoned it not to be so important.

4.2.5 Comprehension

Table 9 shows the only recurring strength discussed by the participants regarding comprehension.

Table 9

Main strength related to the descriptors for comprehension

Strength
Reference to comprehension of cultural subtleties

Some participants commented that one of the main strengths in the descriptors for comprehension is the reference to comprehension of cultural subtleties at level 6. However, participant B argued that it should have been included in other levels as well because, as she argued: "Imagine, if it is difficult for a proficient speaker to deal with cultural differences, imagine for a very low proficient speaker. So it is not only to expert speakers".

Moreover, although two participants pointed out that the explanation about the ability to "read between the lines" at level 6 is helpful, participant B believes that this is not only difficult to test, but inappropriate, as radiotelephony communications need to be clear, concise and unambiguous. She pointed out that "you´ll never expect someone to read between the lines. You need to be clear for everybody to understand you". Although a conclusion cannot be drawn regarding this matter, I personally agree with participant B.

4.2.6 Interactions

Table 10 shows the recurring strength and weaknesses discussed by the participants regarding interactions.

Table 10

Main strength and weaknesses related to the descriptors for interactions

Strength	Weaknesses
Inclusion of ability to check, confirm and clarify	Reference to sensitivity to non-verbal cues
	Level 6 descriptors seem weaker than level 5's

Four of the five participants who answered the questions about interactions believe that one of the main strengths regarding interactions was the inclusion of the ability to check, confirm and clarify. However, two participants argued that this ability should have also been included at levels 5 and 6. Participant C argued that “this idea that checking, confirming and clarifying is a feature that only lower level proficiency users display is wrong, because much higher level listeners also display that feature when there is a situation which is very difficult to comprehend”.

The reference to sensitivity to non-verbal cues was strongly criticised by the participants, as the TLU is the language used in radiotelephony communications, which are always verbal. A number of participants argued that another weakness regarding interactions is that the descriptors for level 6 seem weaker than those for level 5. As pointed out by participant C, “I would rather have somebody that responds immediately appropriately and informatively talking to me than somebody who interacts with ease in nearly all situations, which implies that there are situations where they don’t interact with ease at all”.

4.3 Recurring/relevant themes related to the ICAO LPRs

Besides the issues which have already been discussed, nine other relevant and recurring themes were brought up by the participants during the interviews. It is important to say that the themes, including the previously discussed issues, are related to each other, so they sometimes overlap. Table 11 shows what these themes were.

Table 11

Relevant and recurring themes related to the ICAO LPRs

	Themes		**Sub-themes**
4.3.1	Contradictions in the policy	**4.3.1.1**	Face-to-face communications being included in the holistic descriptors
		4.3.1.2	Policy targeting non-native speakers
4.3.2	Policy does not fit the TLU		
4.3.3	The need to test level 6 candidates’ ability to communicate effectively		

4.3.4	The importance of adhering to standardized phraseology
4.3.5	Participants’ opinions about what rating scale categories they consider to be the most and least important
4.3.6	Rating challenges
4.3.7	Terminologies used in the rating scale are sometimes confusing
4.3.8	The existence of very bad tests in the market and the need for ICAO to take more responsibility towards the LPRs

4.3.1 Contradictions in the policy

All participants criticised the ICAO policy to some extent. Participants argued that the policy contradicts itself. The two main contradictions discussed are related to the difficulty to understanding the reason why ICAO included face-to-face communications in the holistic descriptors, and the fact that the policy targets non-native speakers of English.

4.3.1.1 Face-to-face communications being included in the holistic descriptors

One of the main criticisms regarded the fact that ICAO included face-to-face communications in the holistic descriptors, in spite of the target language being radiotelephony communications. This theme was discussed in four of the interviews. None of the participants understood why ICAO wrote that. Participant C pointed out that face-to-face communications being included in the holistic descriptors “often leads to confusion over what we are testing and how we are testing it”. He went on saying that he thinks “this dilutes the message that we are testing English for a very, very specific purpose, for safe communications. I think it confuses test developers, authorities and test takers. It’s not uncommon for pilots to say why are we doing this? I never do this as part of my job”. Participant A claimed that he knows pilots and ATCs do not meet face-to-face, and commented: “I sometimes ask myself, go back and think what was our intention when we prepared these holistic descriptors, but obviously I was swiped by the academics, the linguists who felt that there was all that value in a face-to-face communication”.

4.3.1.2 Policy targeting non-native speakers

Another recurring sub-theme regarding contradictions in the policy refers to the fact that, despite ICAO's effort to raise awareness that this is an ELF context and their statement that "the ICAO language proficiency requirements apply to native and non-native speakers alike" (ICAO, 2010, 5-4), the policy focuses on non-native speakers. This can be seen in the general features of the policy as well as in the rating scale and the explanations of the descriptors. Speaking of the LPRs in general, participants commented on the fact that native speakers are not required to go through formal assessment and that level 6 candidates do not need to be tested again. Participant B pointed out that "when they don´t require the native-speakers to be formally assessed, they are considering that if they are native speakers, they know how to deal with any problems in radiotelephony communications. But we know that this is not true". She argued:

> So not only native speakers need to be tested, but the test should include skills and competences that they need in order to communicate with non-native speakers of the language. For example, choice of vocab, rate of speech, strategies to accommodate or to clarify things, to be aware of the problems and of the difficulties of the non-native speakers. They also need to be tested in a number of things that are not included in the rating scale. So when you ask me about retesting level 6, first they need to be tested and tested in the correct things, in the correct skills and competencies. And then, of course, they need to be retested. It is not a matter of knowing the language. It is a matter of knowing how to use the language in this context, how to interact appropriately.

The participants also mentioned excerpts from the rating scale and explanations of the descriptors that contradict the ELF theory. For example, the reference to how much the candidates' pronunciation is influenced by the first language in the pronunciation descriptors, the explanation "pronunciation plays the critical role in aiding comprehension between two non-native speakers of English", and the explanation of the descriptors for fluency at level 6 which states that "fluency at this level is native-like or near native-like". Participant C pointed out targeting on non-native speakers is "very unfortunate" and "quite culturally insensitive too". Likewise, participant E believes it to be "insulting".

4.3.2 Policy does not fit the TLU

The majority of participants made comments on features of the policy that do not reflect the TLU domain. To begin with, a number of them expressed dissatisfaction regarding how the ICAO policy was created. They also criticised the fact that there is no information available on the work developed by the PRICESG. Participant B suspected that a thorough needs analysis was not carried out, and participant F believes that some decisions that were taken were not “based on scientific criteria of language testing”. Participant C claimed that ICAO “had very limited time and a tiny budget to get this right, and they didn’t get it right”.

Some participants argued that there are important competencies that need to be assessed that were not included in the rating scale, whereas there are some other skills that were included but are irrelevant to the TLU. For example, participant B argued that “there are some things that are not being taken into account, the strategies, the cultural competence, some authors call it interactional competence. That is necessary (…) so relevant for safety”. She also raised a point that perhaps the candidates’ experience should also be taken into consideration because communication happens as a result of the relationship between a candidate’s language skills and his/her experience. She added that “this context is very complex, it involves a lot of things, not only language. All competences are necessary and I believe they are all part of the language use domain”. However, some participants argued that, although it is important to assess these other skills, it is not easy to design test tasks that will elicit this kind of behaviour.

The skills that were pointed out as probably irrelevant to the TLU include the assessment of idiomatic vocabulary usage, sensitivity to register, sensitivity to non-verbal cues, use of complex structures, use of discourse markers and connectors, and unfamiliar vocabulary. All of the participants advocated that idiomatic vocabulary should not have been included in the assessment criteria. Participant C argued that idiomatic vocabulary “has got no place in radiotelephony communications. It doesn’t necessarily identify strong users from weaker users. It has a deleterious effect on safety and it shouldn’t be there, it has absolutely no place in this rating scale”. Participant B wondered why idioms were included as they “are not part of the target language in radiotelephony communications”. Sensitivity to register and sensitivity to non-verbal cues were also brought up many times. In relation to sensitivity to register, participant

B pointed out that “the context here is unique, you don´t change the context, so the thing about being sensitive or being flexible to register does not make much sense”, whereas participant C claimed “we are talking about one register, and that’s the ability to communicate on the radio. You don’t have multiple registers on the radio. It’s short, brief, concise, to the point, safety operational related language use. There is no room for different registers in that context, so it is nonsense to include it in the scale”. As regards sensitivity to non-verbal cues, it is not clear to most of the participants how ICAO defines them. They argued that anything that is expressed on the radio through the voice is verbal and questioned why sensitivity to non-verbal cues was included. Participant E said that assessing non-verbal cues would be “guesswork”. Similarly, there was also criticism regarding the assessment of complex structures and use of discourse markers and connectors, as pilots and ATCs are expected to keep the communication simple. Assessing unfamiliar topics was also considered to be inappropriate, as participant B argues, it is “something that goes beyond the TLU domain”.

4.3.3 The need to test level 6 candidates’ ability to communicate effectively

The majority of participants believe it to be very important to assess level 6 candidates’ ability to communicate effectively on the radio. The participants advocated that there should be a different test for level 6 candidates because native speakers need to be tested on some skills that are particularly important for them to interact appropriately with non-native speakers, and even with other native speakers. Participants C and D explained that being a native speaker does not mean being an effective interlocutor on the radio. Participant F pointed out that “at level 6 it’s not so much about your language proficiency, but about your communicative ability, and the communicative ability is something that both native and non-native speakers have to learn, probably even more so the native speakers because native speakers rarely think about their language”. As argued by participant B, when talking about native speakers, “many attitudes, many different kinds of behaviour on the radio are influenced by the culture, not only by their national culture, but also by their professional culture, so pilots perform differently from controllers. They have this difference. So if nobody takes this into account, we will have a problem”.

As the participants pointed out, the test tasks would need to be different from those to assess lower levels of proficiency. It is necessary to test if proficient speakers can be cooperative and use accommodation strategies, like using simple structure and vocabulary, paraphrasing, avoiding ambiguity, avoiding idiomatic expressions, adapting their speech rate, improving their pronunciation intelligibility (stress, intonation), in case their interlocutors have a comprehension problem. Participant A highlighted the importance of training native speakers to be always conscious because many times they will be interacting with less proficient interlocutors. Testing these abilities is very important, after all, as participant B argued, native speakers may also fail to communicate.

4.3.4 The importance of adhering to standardized phraseology

Another recurrent theme in the data set was the importance of adhering to standardized phraseology. Three participants mentioned pilots and ATCs often deviate from phraseology and argued that this underutilization of phraseology is one of the main causes of miscommunications. As participant A pointed out, "they want to speak English, they want to say, 'listen to me, my English is very good'. This means that from a professional point of view the phraseology is going down and people want to use more plain language". During the interview he went back to this issue many times. For example, he said that "people want to show how well they can use the language and that´s the danger now, people trying to be too clever with some of the words and phrases they have".

4.3.5 Participants' opinions about what rating scale categories they consider to be the most and the least important

The participants' opinions as regards what rating scale categories are the most and the least important varied considerably. Participants A and E believe they are all equally important, but participant A thinks pronunciation stands out as the most critical area. Participant E argued that perhaps the most important skills are "the ones that affect meaning the most, that would be structure and pronunciation". The four other participants rated them from the most important to the least important. Participant B believes interactions is the most important, followed by comprehension and then pronunciation. She considers structure, vocabulary and fluency to be equally less important. Participant C rated comprehension as the highest in importance "because without your ability to listen and understand what is happening you have no chance of using your spoken language performance to engage in communication". He ranked pronunciation and interactions in second place, vocabulary in third, all of which playing an important role. Next fluency, playing "less of a role" with structure last, "at least as it is captured by the rating scale". Participant D argued that pronunciation, structure and vocabulary are the most important because they form the foundation for the language and the other skills. She claimed fluency, comprehension, and interactions "come out naturally if the base is solid". She rated comprehension and interactions second place, and fluency last. Participant F rated structure and vocabulary as the most relevant, followed by comprehension and then interactions. He considers fluency to be less important and pronunciation, the least. As it can be seen in Table 12, which shows how each participant ranked the categories and their weights, even though the participants' perceptions were different, it seems that pronunciation plays the most important role, whereas fluency seems to be considerably less relevant.

Table 12

Categories rank by participants

P	**Pronunciation**		**Structure**		**Vocabulary**		**Fluency**		**Comprehension**		**Interactions**	
	R	W	R	W	R	W	R	W	R	W	R	W
A	1	6	2	3	2	3	2	3	2	3	2	3
B	3	4	4	2	4	2	4	2	2	5	1	6
C	2	4.5	5	1	3	3	4	2	1	6	2	4.5
D	1	5	1	5	1	5	3	1	2	2.5	2	2.5
E	1	5.5	1	5.5	2	2.5	2	2.5	2	2.5	2	2.5
F	5	1	1	5.5	1	5.5	4	2	2	4	3	3
Sum of weights		**26**		**22**		**21**		**12.5**		**23**		**21.5**
Final rank		**1**		**3**		**5**		**6**		**2**		**4**

Note: P = Participant; R= Rank; W=Weight
The weights were established as follows: 6 points were given to the category ranked first, 5 for the second and so on until the last place, which got 1 point. In case of draws, an average was calculated.

For example: participant E ranked pronunciation and structure first, and the rest second. Thus: 6+5=11/2=5.5 for pronunciation and structure, and 4+3+2+1=10/4=2.5 for the other categories.

4.3.6 Rating challenges

Many participants claimed to have experienced difficulties in rating. First of all, interpreting the rating scale consistently either within one rater (intra-rater reliability) or among a group of raters (inter-rater reliability) was considered to be a challenge. Participant D feels that this is difficult "because the rating scale is not very clear and we may have different interpretations".

The area in which participants seem to have the most difficulty in rating is pronunciation. Five participants have mentioned difficulties in rating this category. Participant F declared that the raters from his testing context struggle with rating pronunciation. He added that he believes pronunciation is assessed very differently around the world as he has never seen a test "where I was completely convinced with the way pronunciation was assessed". This might be because, as he argued, "these are subjective value judgments. What I perceive as being 'rarely' may not be the same thing you think is rarely". Another reason for that, as pointed out by two participants, might be the fact that a lot depends on the raters' background. Participant E suggested raters should be conscious about how much raters' familiarity with the candidates' accent can affect their rating, and they need to listen to their candidates consciously. She also pointed out she has difficulty in rating lisps. Two participants find it difficult to differentiate a level 3 from a 4 in pronunciation. Participant C argued that "level 3 and level 4 is an enormous jump. You can't be a very strong 3 and a very weak 4 and be separated by descriptors which are so widely different". Participant D pointed out difficulties in making a difference between levels 5 and 6 in pronunciation because the two levels are very similarly written.

Participant C has also experienced problems to reliably rate comprehension. He claimed that it would need a "45 minutes or an hour test, so you've got enough items in there, all sufficient levels of difficulty to be able to distinguish with reliability between those levels".

All participants talked about how closely related the categories are. As they are all connected, they argued that it is sometimes difficult to separate them. They especially mentioned the overlap between fluency and interactions, comprehension and interactions, and vocabulary and fluency. Anyhow, as pointed out by participant A, it is not possible to rate all criteria in isolation because we are dealing with language, with communication.

In relation to this topic, participant C made an important comment. He argued that because the rating can be very subjective, it is crucial to set the standard and "stick to it". He pointed out it is sometimes difficult to define the correct approach to follow, but once it is defined, it is important to be consistent with it. If standardizing the rating within one testing context is found to be difficult, standardizing it within a country where different tests are applied, or even around the world is much more problematic. As previously mentioned, ICAO has published the CD and the ICAO RSSTA as an effort to help the rating standardization, but standardizing it worldwide is still an enormous challenge.

4.3.7 Terminologies used in the rating scale are sometimes confusing

A relevant theme which I considered to be important to report concerns the comments made by participant C in relation to the terminology used in the rating scale. He criticised that the scale is "overwordy". For example, it talks about "common, concrete and work-related" topics, in the descriptors for vocabulary and comprehension, "familiar" topics in vocabulary, and "predictable situations" in structure and interactions, "unusual or unexpected situations" in structure, "linguistic or situational complication" in comprehension, and "unexpected turn of events" in interactions. Participant C also argued that "situational complications and linguistic complications co-occur, so they happen at the same time, one doesn't happen separately from the other". The use of different terms to describe similar or closely related concepts can be confusing. He recommended that scale writers should choose the best term according to what they mean instead of "using a variety of ways of describing that same thing".

4.3.8 The existence of bad quality tests in the market and the need for ICAO to take more responsibility towards the LPRs

A number of participants expressed concern in relation to the quality of some of the tests that have been used to assess pilots and ATCs' English proficiency in accordance with the ICAO LPRs. They mentioned the following problems: some tests consist of mostly phraseology; at some testing providers it is possible "to buy" a level, and; some tests have items that are too technical. Participant F pointed out that "there are still a lot of very bad tests out there. Really, really dangerous tests, unprofessional tests, unscientific tests, and tests that simply don't work". Be that as it may, I believe the test service providers are not the only ones to be blamed for. As discussed here, the policy is sometimes unclear and needs improvement.

Participant F discussed an interesting point. He talked about the importance of getting ICAO to be more involved with the implementation of the LPRs. He argued that ICAO should make more effort to strengthen the LPR provisions, instead of trying to put the responsibility solely on the States. As he narrated:

> There was this big meeting in Montreal last year or two years ago which was quite funny because ICAO representatives were sitting there and they were saying "our aim is to work ourselves out of the job. The language proficiency requirements have to go into implementation now. It is your job to implement this". It became quite clear at that meeting that this is not going to happen. ICAO still has a very long way to go with this.

It is known that ICAO works with limited personnel and a tight budget. However, I consider it to be very important for the organization to, as the SARPs developer, at least revise and improve the policy. As an extension, the quality of the tests will rise, and safety will be enhanced.

5. Discussion of results

In chapter 5, I relate my findings to the studies mentioned in the literature review, and talk about the conclusions that can be drawn from my findings.

5.1 How the findings relate to the studies mentioned in the literature review

Most results from the studies discussed in chapter 2 were upheld by this research. They were: criticism regarding the policy targeting on non-native speakers; the need to research the nature of aviation English; the importance of adhering to phraseology; the need to define the test construct better; the fact that the policy does not reflect the TLU domain; the need to train and assess interactional competence (including awareness of intercultural factors); the need for rating scale validation work; SME raters have difficulty in using the rating scale; the need to investigate if level 4 is enough for safe international flying; the difficulty to separate the categories; the importance of standardizing the approach in each testing context; the need to revise the rating scale; criticism towards the glossary of basic and complex structures; the need to clarify some of the terminologies used in the scale; how much the assessment of pronunciation depends on the background of the rater; and the need for ICAO to take a more active role in the implementations of the requirements. Another finding that corroborates results from the previous research discussed in chapter 2 is that among the six rating scale criteria, the category which appears to be the most important is pronunciation. The results also upheld Kim and Elder (2009)'s finding, instead of Knoch (2009)'s, that fluency seems to play a less important role. However, it cannot be concluded by this research that structure is also less relevant, as Knoch, and Kim and Elder suspected, because, according to the results of this study, structure seems to be similarly important as the three remaining categories.

In relation to the subject of disregarding technical knowledge of operations when rating, this research has confirmed previous findings which indicated SMEs have difficulties in separating language proficiency from technical knowledge (Davies, 2001; Ryan, 2007; Knoch, 2010). However, most participants, as discussed in chapter 4, section 4.1, agreed with ICAO that raters should not be influenced by candidates' lack of knowledge of operations. They argued that technical knowledge is assessed at a different moment by either a flight examiner, a simulator instructor or a theoretical knowledge test, and that the purpose of the ICAO test is to assess the candidate's ability to speak and understand plain English. I think this difficulty faced by SMEs is related to confidence (lack of confidence is mentioned twice in the explanations for level 3). In my opinion, SME raters might not feel confident when they realize that a candidate lacks operational knowledge and that interferes in their rating of the candidate's language proficiency. The participants suggested that the operational raters should be trained to assess language ability without being negatively influenced by candidates' lack of background knowledge. However, as discussed in chapter 2, and as argued by participant C, it is undeniable that there is an intimate link between the assessment of background knowledge and the assessment of language proficiency in this context (Douglas, 2014; Emery, 2014; Knoch, 2009).

5.2 Main conclusions that can be drawn from my findings

Many interesting conclusions can be drawn from the findings. Some of them are related to the policy in general, and others refer to the rating criteria. One of the main conclusions is that although phraseology should be tested separately from plain English, it is crucial for it to be tested, as its underutilization contributes significantly to communication failures. Unfortunately, English phraseology is not formally tested in some countries (or it is tested only once). I consider it to be very important for ICAO to develop similar guidelines regarding the assessment of phraseology. Even though the results indicated that it would probably be better to test phraseology separately from plain English, it is very important to include phraseology in the language proficiency test, as some participants argued and as advocated by Moder and Halleck (2009), including how candidates deal with the important but difficult to see transition from phraseology to plain English. As suggested by many participants, in case weakness in either the wrong use of phraseology or lack of technical knowledge of operations is noted, it is important to take some kind of action.

In regard to the intervals of assessment proposed by ICAO, there is need for research to investigate "how closely the policy aligns with actual language decay" as this kind of decision must "be borne out by evidence" (participant C). As discussed by the participants, these intervals seem to be too long, as candidates might not be operational after three years in case of a level 4, or six years in case of a level 5. The participants suggested these intervals should be reduced to two years for level 4s and probably four years for level 5s. Moreover, level 6s should definitely be retested.

Another important conclusion is that native or native-like speakers of English need to be formally tested as they need to demonstrate their ability to communicate effectively on the radio. Assessing interactional competence, including awareness of cultural factors, is essential. There should be a specific test to assess highly proficient pilots and ATC's communicative abilities, but these skills are also vital to less proficient speakers.

A very important conclusion that can be drawn from this research is that the rating criteria need to be revised in order to better reflect the TLU domain. First of all, face-to-face communications should be excluded from the holistic descriptors (at least until a better definition of the construct is developed). Face-to-face communications happen in tests which are designed according to the broad interpretation of the test tasks, as previously discussed in this dissertation. However, if the ICAO LPRs target exclusively the language used in pilots/ATCs communication, face-to-face interactions are irrelevant. Also, participant B criticised the fact that the policy allows for these two different interpretations and argued that "this not only makes a difference, but we have a lack of standards in different tests if everything is acceptable".

Regarding the rating scale more specifically, there are important skills that need to be tested which were not included in it, and, on the other hand, skills that go against the TLU were included. Thus, it is crucial to develop a rating scale which reflects the real-life situation. However, for that to happen, it is necessary, first, to better understand the nature of the English used by pilots and ATCs' in radiotelephony communications. Research will assist the development of a definition of the appropriate test construct. It seems to be necessary to include comprehension of cultural factors in the descriptors for levels 4 and 5, as well as ability to check, confirm and clarify in levels 5 and 6. Communicative abilities should be better captured in the rating scale. Furthermore, it is important to clarify the terminologies used in the rating scale and use them consistently. As pointed out by Douglas (2004), use of unclear terminologies is a potential threat to validity. Moreover, the development of a scale dedicated to comprehension might be beneficial. Having a different rating scale to be used by ELEs and SMEs should also be considered. All the results mentioned in this paragraph confirmed findings from studies discussed in chapter 2. Table 13 lists the aspects of the rating scale which should be maintained, deleted, changed, and further investigated. The findings marked with a star (*) corroborate the results from studies debated in the literature review.

Table 13

What should be maintained, deleted from or changed in the rating criteria

	Should be maintained	**Should be deleted**	**Should be changed**	**Should be researched**
Pronunciation	Focus on comprehensibility*	Any reference to influence by the first language* The part "between two non-native speakers of English" should be deleted from the explanation "pronunciation plays the critical role in aiding comprehension between two non-native speakers of English"*	The interference with ease of understanding should not be differentiated only in terms of frequency*	
Structure	Focus on interference with meaning	Reference to complex structures, as pilots and ATCs should not use complex structures*		What aspects interfere with meaning*
Vocabulary	Reference to ability to paraphrase	Reference to idiomatic vocabulary* Reference to sensitivity to register* Reference to unfamiliar topics*		
Fluency		Reference to discourse markers and connectors* Reference to native-like fluency*		Impact of fluency on safety* Recommended rate of 100 words per minute

				Importance of varying speech flow for stylistic device*
Comprehension	Reference to comprehension of cultural subtleties, which should be included in levels 4 and 5*			Importance of ability to "read between the lines"
				How to assess comprehension of cultural subtleties*
				How to rate accents from listening tasks in terms of how "sufficiently intelligible" they are*
Interactions	Reference to ability to check, confirm and clarify, which should be included in levels 5 and 6*	Reference to sensitivity to non-verbal cues*	Levels 5 and 6 descriptors should be revised, as level 6 seems weaker than level 5*	

It is very important to better understand the nature of pilots/ATCs' communications. Only with a clear understanding, the test construct will reflect the language abilities and strategic competence needed in the TLU domain. The rating scale should be developed by empirically based methods and "the categories included in a rating scale should be based on a theory of language, language development or language use" (Knoch, 2009, p. 22). Jacoby and McNamara (1999) argued that rating scales should be developed through an understanding of what it means to know and use a language in the specific domain. Participant C also suggested that "the rating scale revision" should be "based on actual language use rather than prescriptive language use". In case ICAO decides to revise the LPRs, it is highly recommended for them to publish detailed information about all phases of the process, especially information about the validation of the scale/scales. Douglas (2001), Jacoby and McNamara, Kim and Elder (2009), and Knoch (2009) argued that validation research should involve more indigenous assessment, as LSP testing relies on a combination of both professional and language testing expertise. Moreover, Farris et al (2008) advised that it is important to validate the scale “using a large population of pilots and controllers under conditions of varying workload or psychological stress typical of the controller-pilot workplace” (p. 407). They believe that is important because their findings showed pilots language proficiency in air traffic communication was affected by workload. The higher the workload, the more affected their proficiency was. Less proficient participants’ proficiency was more affected by high cognitive workload. Finally, I think it is important for the ICAO LPRs future revision group to develop strategies to ensure raters interpret the descriptors similarly and apply them consistently.

6. Conclusions and implications for future research

This last chapter includes the limitations of this study and the ethical issues involved in it, suggestions for further research, as well as a final conclusion of what the research as a whole has shown.

6.1 Limitations and ethical issues

Some limitations of the study should be acknowledged. Firstly, because of the nature of qualitative research, the results cannot be generalised. Secondly, there was only one SME among the participants, so the results reflect more ELEs' perceptions than SMEs'. Thirdly, although I made an effort to write the questions in a neutral way and to conduct the interview neutrally in order not to influence participants' responses, I might have unintentionally expressed bias when approaching some topics. Another limitation relates to the fact that the interviews were conducted on Skype, so I did not get the benefits of face-to-face communications. More importantly, the generated data need to be treated as an indirect reflection of the participants' understanding of the issues, as I gathered information about what the participants claim they think, which might be different from what they actually think. Finally, the results I presented show my subjective interpretation of the data, which might have been biased by my expectations.

There are no major issues related to ethics regarding the present study. However, it is important to note that all participants were sent a participant information sheet (see Appendix I) and were asked to sign a consent form (see Appendix J). The participants were guaranteed that the data would be anonymized, although participant C said he would not mind being identified. I would like to point out that deciding what information to include about the participants in the methodology chapter was difficult because, at the same time that I needed to explain why the participants were interesting and unique test developers and raters, I needed to guarantee anonymity.

6.2 Suggestions for further research

There is need for further research beyond the areas which were pointed out in Table 13. The results from this study corroborate the idea that it is important to carry out extensive research to better understand the nature of the language used in pilots/ATCs' communications (Douglas, 2004; Emery, 2014; Moder & Halleck, 2009). The studies should investigate if it is also important to assess reading and writing. It is equally important to find out if the TLU should include other types of communication (i.e. pilot/pilot, pilot/mechanic, ATC/supervisor communications) or if other tests should be developed to assess that. Another interesting topic for research would be to investigate if level 4 is enough for safe international operations, as suggested by participant C:

> I think we really need to understand today how pilots and controllers who today by and large have an ICAO level 4 plus perceive the effectiveness of communication. For example, today if you ask pilots or controllers whether the people they talk to on the radio are good enough to do the job, this will tell us whether ICAO level 4 is functioning as intended in terms of establishing a minimum level of proficiency for safe communications.

Furthermore, research should be conducted to investigate what the appropriate intervals for reassessment are, as this kind of decision must be supported by evidence. It is essential to do more research to further investigate the current rating criteria. Knowing the strengths and weaknesses of the descriptors can help the development of a better rating scale. I also think it is important to investigate the reasons why previous research has indicated low reliability coefficients (Pfeiffer, 2009; Garcia, 2014). Rating inconsistency is likely to be linked to deficiencies in the rating scale. Thus, it is important to know what aspects of the scale are leading raters to different interpretations. As previously discussed, it is necessary to conduct research to find out if it is interesting to develop an exclusive rating scale for comprehension, and, in case it is, what aspects it should include. Additionally, as the results confirmed Knoch (2009)'s finding that the current scale seems to be difficult to be used by operational raters, the relevance of developing a specific rating scale to be used by SMEs should be researched. As Knoch (2014) hypothesised: "it is possible that industry specialists are best asked to make simple yes/no judgements rather than using complicated scale criteria" (p.85). Another kind of research that would help to collect relevant information about how to improve the LPRs would be to investigate the strategies test developers and raters have developed to deal with the challenges they face. As recommended by Kim and Elder (2009), research should also be undertaken in non-English speaking countries.

6.3 Concluding remarks

As argued by participant A, with the implementation of the ICAO LPRs “the level of English in general terms has come up considerably in communication”, but “it will never be perfect”. Although it will never be perfect, the more we work on developing professional standards, the more we improve safety. As seen in the previous section, there is still a lot to be discussed. As argued by Read and Knoch (2009), “the whole topic of oral communication in the aviation context is likely to engage the attention of language testers and other applied linguists for some time to come” (p. 21.10). Nevertheless, I urge ICAO to consider the results of this and other studies and to take actions towards the establishment of a group to revise the LPRs. As argued by two participants, ICAO will only revise the LPRs if a safety case is built. As participant C argued, “having an unreliable scale is enough of a safety case”. My overall conclusion is that, although the ICAO LPRs have been a remarkable advance, twelve years have passed and the time has come to revise them. Although it is never going to be perfect, the policy can and should be improved.

References

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Alderson, J. C. (2009). Air safety, language assessment policy, and policy implementation: the case of aviation English. *Annual Review of Applied Linguistics*, 29, 168-187. doi:10.1017/S0267190509090138

Alderson, J. C. (2010). A survey of Aviation English Tests. *Language Testing*, 27(1), 51-72. doi: 10.1177/0265532209347196

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

Braun, V. & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101.

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

Cohen, L., Morrison, K., & Manion, L. (2007). *Research Methods in Education*. London: Routledge.

Davies, A. (2001). The logic of testing languages for specific purposes. *Language Testing*, 18(2), 133–147.

Dörnyei, Z. (2007). *Research methods in applied linguistics: quantitative, qualitative and mixed methodologies*. 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), 171–185.

Douglas, D. (2004). Assessing the language of international civil aviation: Issues of validity and impact. Proceedings from the *International Professional Communication Conference*, IEEE Professional Communication Society (pp. 248-252). Minneapolis: IEEE.

Douglas, D. (2014). Nobody seems to speak English here today: Enhancing assessment and training in aviation English. *Iranian Journal of Language Teaching Research, 2*(2), 1-12.

Emery, H. J. (2014). Developments in LSP testing 30 years on? The case of aviation English. *Language Assessment Quarterly*, 11:2, 198-215. doi: 10.1080/15434303.2014.894516

Foy, V. (2012, November). [*No title*]. Plenary speech presented at the International Civil Aviation English Association (ICAEA) Forum. Bangkok, Thailand.

Garcia, A. C. M. (2014). *Testing English for the specific purpose of aeronautical communications: issues in pronunciation assessment*. (Unpublished Masters module assignment). Lancaster University, Lancaster, UK.

Farris, C., Trofimovich, N., Segalowitz, N., & Gatbonton, E. (2008). Air traffic communication in a second language: implications of cognitive factors for training and assessment. *TESOL Quarterly*, 42(3), 397–410.

Harding, L. (2014, May). How can we know if pronunciation techniques and activities have been successful? Assessing pronunciation in the classroom and beyond. *TESOL Education programs*. Seminar conducted from TESOL Virtual.

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

Huhta, A. (2009). An analysis of the quality of English testing for aviation purposes in Finland. *Australian Review of Applied Linguistics*, 32(3), 26.1–26.14. doi: 10.2104/aral0926.

Jacoby, S., & McNamara, T. (1999). Locating competence. *English for Specific Purposes*, 18(3), 213–241.

Jenkins, J. (2000). *The phonology of English as an international language*. Oxford, England: Oxford University Press.

International Civil Aviation Organization. (2010). *Manual on the Implementation of ICAO Language Proficiency Requirements, 2nd Edition* (Doc. 9835). Montreal, Canada: Author.

Kim, H., & Elder, C. (2009). Understanding Aviation English as a lingua franca: Perceptions of Korean aviation personnel. *Australian Review of Applied Linguistics*, 32 (3), 23.1-23.17.

Kim, H. (2013). Exploring the construct of radiotelephony communication: A critique of the ICAO English testing policy from the perspective of Korean aviation experts. *Papers in Language Testing and Assessment*, *2*(2), 103-110.

Kim, H., & Elder, C. (2014). Interrogating the construct of aviation English: Feedback from test takers in Korea. *Language Testing*, 0265532214544394.

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

Knoch, U. (2014). Using subject specialists to validate an ESP rating scale: The case of the International Civil Aviation Organization (ICAO) rating scale. *English for Specific Purposes*, *33*, 77-86.

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

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

Moder, C. L., & Halleck, G. B. (2009). Planes, politics and oral proficiency. *Australian Review of Applied Linguistics*, 32(3), 25.1–25.16. doi: 10.2104/aral0925.

Monteiro, A. L. T. (2012). Radiotelephony communications: threats in a multicultural context. *Aviation in Focus,* 3(2), 44-66.

Morrow, D., Rodvold, M., & Lee, A. (1994). Nonroutine transactions in controller-pilot communication. *Discourse Processes*, 17, 235-258.

Polkinghorne, D. E. (2005). Language and meaning: data collection in qualitative research. *Journal of Counseling Psychology*, 52/2, 137-145.

Pfeiffer, A. (2009). *Inter-rater reliability in an aviation speaking test*. Unpublished Masters dissertation. Lancaster University, Lancaster, UK.

Prado, M. C. A. (2015). *Levantamento de padrões léxico-gramaticais do inglês para aviação: um estudo vetorado pela Linguística de Corpus.* [Survey on the lexico-grammatical patterns in aviation English: a study conducted by corpus linguistics]. Unpublished Masters dissertation, Universidade de São Paulo, São Paulo, Brazil.

Prinzo, V. (2009) *The ICAO English language proficiency rating scale applied to enroute voice communications of U.S. and foreign pilots*. DOT/FAA/AM-09/10, Washington D.C., FAA publishing.

Read, J., & Knoch, U. (2009). Clearing the air: Applied linguistic perspectives on aviation communication. *Australian Review of Applied Linguistics*, 32(3), 21.1–21.11. doi: 10.2104/aral0921.

Ryan, K. (2007). Assessing the OET: *The nurse's perspective. Unpublished Masters thesis*. The University of Melbourne, Melbourne.

Scaramucci, M. (2011). O exame de proficiência em língua inglesa para controladores de voo do SISCEAB: Uma entrevista com Matilde Scaramucci. *Aviation in Focus, 2* (1), 3-12.

Tiewtrakul, T., & Fletcher, S. R. (2010). The challenge of regional accents for aviation English language proficiency standards: A study of difficulties in understanding in air traffic control pilot communication. *Ergonomics*, 53(2), 229–239.

Widdowson, H. G. (1994). The ownership of English. *TESOL Quarterly*, 28(2), 377–389.

Appendix A: the holistic descriptors

Proficient speakers shall:

a) communicate effectively in voice-only (telephone/radiotelephone) and in face-to-face situations;

b) communicate on common, concrete and work-related topics with accuracy and clarity;

c) use appropriate communicative strategies to exchange messages and to recognize and resolve misunderstandings (e.g. to check, confirm, or clarify information) in a general or work-related context;

d) 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 work situation or communicative task with which they are otherwise familiar; and

e) use a dialect or accent which is intelligible to the aeronautical community.

Appendix B: the 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	Pronunciation, stress, rhythm, and intonation, though possibly influenced by the first language or regional variation, almost never interfere with ease of understanding.	Both basic and complex grammatical structures and sentence patterns are consistently well controlled.	Vocabulary range and accuracy are sufficient to communicate effectively on a wide variety of familiar and unfamiliar topics. Vocabulary is idiomatic, nuanced, and sensitive to register.	Able to speak at length with a natural, effortless flow. Varies speech flow for stylistic effect, e.g. to emphasize a point. Uses appropriate discourse markers and connectors spontaneously.	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 ease of understanding.	Basic grammatical structures and sentence patterns are consistently well controlled. Complex structures are attempted but with errors which sometimes interfere with meaning.	Vocabulary range and accuracy are sufficient to communicate effectively on common, concrete, and work-related topics. Paraphrases consistently and successfully. Vocabulary is sometimes idiomatic.	Able to speak at length with relative ease on familiar topics but may not vary speech flow as a stylistic device. Can make use of appropriate discourse markers or connectors.	Comprehension 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 events. 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 4	Pronunciation, stress, rhythm, and intonation are influenced by the first language or regional variation but only sometimes interfere with ease of understanding.	Basic grammatical structures and sentence patterns are used creatively and are usually well controlled. Errors may occur, particularly in unusual or unexpected circumstances, but rarely interfere with meaning.	Vocabulary range and accuracy are usually sufficient to communicate effectively on common, concrete, and work-related topics. Can often paraphrase successfully when lacking vocabulary in unusual or unexpected circumstances.	Produces stretches of language at an appropriate tempo. There may be occasional loss of fluency on transition from rehearsed or formulaic speech to spontaneous interaction, but this does not prevent effective communication. Can make limited use of discourse markers or connectors. Fillers are not distracting.	Comprehension 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 apparent misunderstandings by checking, confirming, or clarifying.
Levels 1, 2 and 3 are on subsequent page.						

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*
Levels 4, 5 and 6 are on preceding page.						
Pre-operational 3	Pronunciation, stress, rhythm, and intonation are influenced by the first language or regional variation and frequently interfere with ease of understanding.	Basic grammatical structures and sentence patterns associated with predictable situations are not always well controlled. Errors frequently interfere with meaning.	Vocabulary range and accuracy are often sufficient to communicate on common, concrete, or work-related topics, but range is limited and the word choice often inappropriate. Is often unable to paraphrase successfully when lacking vocabulary.	Produces stretches of language, but phrasing and pausing are often inappropriate. Hesitations or slowness in language processing may prevent effective communication. Fillers are sometimes distracting.	Comprehension is often accurate on common, concrete, and work- related 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.	Responses are sometimes immediate, appropriate, and informative. Can initiate and maintain exchanges with reasonable ease on familiar topics and in predictable situations. Generally inadequate when dealing with an unexpected turn
Elementary 2	Pronunciation, stress, rhythm, and intonation are heavily influenced by the first language or regional variation and usually interfere with ease of understanding.	Shows only limited control of a few simple memorized grammatical structures and sentence patterns.	Limited vocabulary range 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.

Appendix C: Explanation of rating scale descriptors (level 3 and above)

Pronunciation

The six levels of pronunciation descriptors are applicable at all levels to native and non-native speakers. This implies that native English speakers may demonstrate Elementary Level 2 proficiency if their regional dialect is so localized that it is not readily understood by those outside of that particular region. On the other hand, speakers whose speech patterns clearly identify them as non-native speakers (having a so-called "accent") may demonstrate Expert Level 6 proficiency, as long as this meets the criterion of "almost never" interfering with ease of understanding.

Pre-operational 3: *Pronunciation, stress, rhythm and intonation are influenced by the first language or regional variation and frequently interfere with ease of understanding.*	***Operational 4:*** *Pronunciation, stress, rhythm and intonation are influenced by the first language or regional variation, but only sometimes interfere with ease of understanding.*	***Extended 5:*** *Pronunciation, stress, rhythm and intonation, though influenced by the first language or regional variation, rarely interfere with ease of understanding.*	***Expert 6:*** *Pronunciation, stress, rhythm and intonation, though possibly influenced by the first language or regional variation, almost never interfere with ease of understanding.*
Accent at this Pre-operational Level 3 is so strong as to render comprehension by an international community of aeronautical radiotelephony users very difficult or impossible. It should be noted that native or second language speakers may be assessed at this level in cases where a regional variety of the language has not been sufficiently attenuated.	Operational Level 4 speakers demonstrate a marked accent, or localized regional variety of English. Occasionally, a proficient listener may have to pay close attention to understand or may have to clarify something from time to time. Operational Level 4 is certainly not a perfect level of proficiency; it is the minimum level of proficiency determined to be safe for air traffic control communications. While it is not an Expert level, it is important to keep in mind that pronunciation plays the critical role in aiding comprehension between two non-native speakers of English.	Extended Level 5 speakers demonstrate a marked accent, or localized regional variety of English, but one which rarely interferes with how easily understood their speech is. They are always clear and understandable, although, only occasionally, a proficient listener may have to pay close attention.	An Expert Level 6 speaker may be a speaker of English as a first language with a widely understood dialect or may be a very proficient second-language speaker, again with a widely used or understood accent and/or dialect. The speakers' accent or dialect may or may not identify them as second language users, but the pronunciation patterns or any difficulties or "mistakes" almost never interfere with the ease with which they are understood. Expert speakers are always clear and understandable.

Structure

Relevant grammatical structures and sentence patterns are determined by language functions appropriate to the task. Users may refer to the communicative aeronautical language functions, to the list of controller communicative tasks and to the classification of basic and complex structures in Appendix B for guidance. Language teaching specialists generally categorize grammatical errors into two classes: "global" and "local". Global errors are those which interfere with meaning; local errors are those which do not interfere with meaning.

Pre-operational 3: Basic grammatical structures and sentence patterns associated with predictable situations are not always well controlled. Errors frequently interfere with meaning.	*Operational 4: 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.*	*Extended 5: Basic grammatical structures and sentence patterns are consistently well controlled. Complex structures are attempted but with errors which sometimes interferes with meaning.*	*Expert 6: Both basic and complex grammatical structures and sentence patterns are consistently well controlled.*
A weak command of basic grammatical structures at this level will limit available range of expression or result in errors which could lead to misunderstandings.	Operational Level 4 speakers have good command of basic grammatical structures. They do not merely have a memorized set of words or phrases on which they rely but have sufficient command of basic grammar to create new meaning as appropriate. They demonstrate local errors and infrequent global errors and communication is effective overall. Level 4 speakers will not usually attempt complex structures, and when they do, quite a lot of errors would be expected resulting in less effective communication.	Extended Level 5 speakers demonstrate greater control of complex grammatical structures than do Operational Level 4 speakers and may commit global errors from time to time when using complex structures. The critical difference between the Level 4 and Level 5 requirements concerns the use of basic grammatical structures and sentence patterns compared to the use of complex structures (see the glossary of basic and complex structures in Appendix B, Part IV). At Level 5, the structure descriptors refer to consistent control of basic structure, with errors possibly occurring when complex structures and language are used. There is actually a big jump between Level 4 and Level 5. Level 5 speakers will have a more sophisticated use of English overall, but will exhibit some errors in their use of complex language structures, but not in their basic structure patterns.	Expert Level 6 speakers do not demonstrate consistent global structural or grammatical errors but may exhibit some local errors.

Vocabulary

Vocabulary includes individual words and fixed expression. Vocabulary can be classified by the domains of meaning to which it refers. A partial list of vocabulary domains related to aviation communications is found in Appendix B of this manual. While memorizing phraseologies is neither an acceptable means of demonstrating language proficiency nor an effective or recommended language learning strategy, it is undeniable that *context* is a relevant factor in language proficiency. Therefore, learning or testing that focuses on, or is designed to elicit vocabulary related to, aeronautical radiotelephony communications is preferable.

Pre-operational 3: *Vocabulary range and accuracy are often sufficient to communicate on common, concrete or work-related topics, but range is limited and the word choice often inappropriate. Is often unable to paraphrase successfully when lacking vocabulary.*	***Operational 4:*** *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.*	***Extended 5:*** *Vocabulary range and accuracy are sufficient to communicate effectively on common, concrete and work-related topics. Paraphrases consistently and successfully. Vocabulary is sometimes idiomatic.*	***Expert 6:*** *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.*
Gaps in vocabulary knowledge and/or choice of wrong or non-existent words are apparent at this level. This has a negative impact on fluency or results in errors which could lead to misunderstandings. The frequent inability to paraphrase unknown words or in the process of clarification makes accurate communication impossible.	An Operational Level 4 speaker will likely not have a well-developed sensitivity to register (see glossary on page (ix)). A speaker at this level will usually be able to manage communication on work-related topics, but may sometimes need clarification. When faced with a communication breakdown, an Operational Level 4 speaker can paraphrase and negotiate meaning so that the message is understood. The ability to paraphrase includes appropriate choices of simple vocabulary and considerate use of speech rate and pronunciation.	Extended Level 5 speakers may display some sensitivity to register, with a lexical range which may not be sufficient to communicate effectively in as broad a range of topics as an Expert Level 6 speaker, but a speaker with Extended proficiency will have no trouble paraphrasing whenever necessary.	Level 6 speakers demonstrate a strong sensitivity to register. Another marker of strong proficiency seems to be the acquisition of, and facility with, idiomatic expressions and the ability to communicate nuanced ideas. As such, use of idioms may be taken into account in assessment procedures designed to identify Level 6 users in a non-radiotelephony context. This is not however intended to imply that idiomatic usages are a desirable feature of aeronautical radiotelephony communications. On the contrary, use of idioms is an obstacle to intelligibility and mutual understanding between non-expert users and should therefore be avoided by all users in this environment.

Fluency

For our purposes, fluency is intended to refer to the naturalness of the flow of speech production, the degree to which comprehension is hindered by any unnatural or unusual hesitancy, distracting starts and stops, distracting fillers (em … huh … er …) or inappropriate silence. Levels of fluency will be most apparent during longer utterances in an interaction. They will also be affected by the degree of expectedness of the preceding input which is dependent on familiarity with scripts or schemata described in Chapter 3.

Pre-operational 3: *Produces stretches of language, but phrasing and pausing are often inappropriate. Hesitations or slowness in language processing may prevent effective communication. Fillers are sometimes distracting.*	***Operational 4:*** *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.*	***Extended 5:*** *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.*	***Expert 6:*** *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.*
The slowness of speech flow at this level is such that communication lacks concision and efficiency. Long silent pauses frequently interrupt the speech flow. Speakers at this level will fail to obtain the professional confidence of their interlocutors.	Speech rate at this level may be slowed by the requirements of language processing, but remains fairly constant and does not negatively affect the speaker's involvement in communication. The speaker has the possibility of speaking a little faster than the ICAO recommended rate of 100 words per minute if the situation requires (Annex 10, Volume II, 5.2.1.5.3 b)).	Rate of speech and organization of discourse at this level approach natural fluency. Under appropriate circumstances, rates significantly higher than the ICAO recommended rate of 100 words per minute can be achieved without negatively affecting intelligibility.	Fluency at this level is native-like or near native-like. It is notably characterized by a high degree of flexibility in producing language and in adapting the speech rate to the context of communication and the purposes of the speaker.

Comprehension

This skill refers to the ability to listen and understand. In air traffic control communications, pilots rely on the clear and accurate information provided to them by controllers for safety. It is not sufficient for air traffic controllers to be able to handle most pilot communications; they must be ready for the unexpected. Similarly, pilots must be able to understand air traffic controller instructions, especially when these differ from what a pilot expects to hear. It is during complications in aviation that communications become most crucial, with a greater reliance upon plain language. While comprehension is only one out of six skills in the Rating Scale, it represents half of the linguistic workload in spoken communications.

Pre-operational 3: *Comprehension is often accurate on common, concrete and work-related 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.*	***Operational 4:*** *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.*	***Extended 5:*** *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 events. Is able to comprehend a range of speech varieties (dialect and/or accent) or registers.*	***Expert 6:*** *Comprehension is consistently accurate in nearly all contexts and includes comprehension of linguistic and cultural subtleties.*
Level 3 comprehension is limited to routine communications in optimum conditions. A pilot or controller at this level would not be proficient enough to understand the full range of radiotelephony communications, including unexpected events, substandard speech behaviours or inferior radio reception.	As with all Operational Level 4 descriptors, comprehension is not expected to be perfectly accurate in all instances. However, pilots or air traffic controllers will need to have strategies available which allow them to ultimately comprehend the unexpected or unusual communication. Unmarked or complex textual relations are occasionally misunderstood or missed. The descriptor of Operational Level 4 under "Interactions" clarifies the need for clarification strategies. Failure to understand a clearly communicated unexpected communication, even after seeking clarification, should result in the assignment of a lower proficiency level assessment.	Level 5 users achieve a high degree of detailed accuracy in their understanding of aeronautical radiotelephony communications. Their understanding is not hindered by the most frequently encountered non-standard dialects or regional accents, nor by the less well-structured messages that are associated with unexpected or stressful events.	Level 6 users achieve a high degree of detailed accuracy and flexibility in their understanding of aeronautical radiotelephony communications regardless of the situation or dialect used. They further have the ability to discern a meaning which is not made obvious or explicit ("read between the lines"), using tones of voice, choice of register, etc., as clues to unexpressed meanings.

Interactions

Because radiotelephony communications take place in a busy environment, the communications of air traffic controllers and pilots must not only be clear, concise and unambiguous, but appropriate responses must be delivered efficiently and a rapid response time is expected. The interactions skill refers to this ability, as well as to the ability to initiate exchanges and to identify and clear up misunderstandings.

Pre-operational 3: *Responses are sometimes immediate, appropriate and informative. Can initiate and maintain exchanges with reasonable ease on familiar topics and in predictable situations. Generally inadequate when dealing with an unexpected turn of events.*	***Operational 4:*** *Responses are usually immediate, appropriate and informative. Initiates and maintains exchanges even when dealing with an unexpected turn of events. Deals adequately with apparent misunderstandings by checking, confirming or clarifying.*	***Extended 5:*** *Responses are immediate, appropriate and informative. Manages the speaker/listener relationship effectively.*	***Expert 6****: Interacts with ease in nearly all situations. Is sensitive to verbal and nonverbal cues and responds to them appropriately.*
The interaction features at this level are such that communication lacks concision and efficiency. Misunderstandings and non-understandings are frequent leading to possible breakdowns in communication. Speakers at this level will not gain the confidence of their interlocutors.	A pilot or air traffic controller who does not understand an unexpected communication must be able to communicate that fact. It is much safer to query a communication, to clarify, or even to simply acknowledge that one does not understand rather than to allow silence to mistakenly represent comprehension. At Operational Level 4, it is Acceptable that comprehension is not perfect 100 per cent of the time when dealing with unexpected situations, but Level 4 speakers need to be skilled at checking, seeking confirmation, or clarifying a situation or communication.	Interactions at this level are based on high levels of comprehension and fluency. While skills in checking, seeking confirmation and clarification remain important, they are less frequently deployed. On the other hand speakers at this level are capable of exercising greater control over the conduct and direction of the conversation.	Expert speakers display no difficulties in reacting or initiating interaction. They are additionally able to recognize and to use non-verbal signs of mental and emotional states (for example, intonations or unusual stress patterns). They display authority in the conduct of the conversation.

Appendix D: Summary of how the rating scale deals with each category

Pronunciation: The ICAO rating scale descriptors for pronunciation can be divided into two different parts. The first one refers to how much pronunciation, stress, rhythm and intonation are influenced by the first language or regional variation (level 2's is heavily influenced, while level 6's says "though possibly influenced"), whereas the second part refers to how frequently they interfere with ease of understanding (level 2's "usually interfere with ease of understanding" and level 6's "almost never interfere").

Structure: The descriptors for structure encompass control of basic structures, which are usually well controlled by level 4 candidates, and ability to use complex structure, which are attempted by level 5 candidates and are well controlled by level 6 candidates. The type and frequency of errors are also important to define a candidate's level in structure. ICAO talks about two types of errors: errors which interfere with meaning, so called global errors; and errors which do not interfere with meaning, the local errors (ICAO, 2010, p. 4-10). In the second edition of the DOC 9836, ICAO, in Appendix B, Part IV, ICAO publish a glossary of what may be considered basic and complex structures, which had not been defined in the first edition of the document (see Appendix D).

Vocabulary: The vocabulary descriptors stipulate if vocabulary range and accuracy are sufficient to communicate on common, concrete and work-related topics, wherein level 6 candidates' vocabulary range and accuracy are sufficient to communicate effectively on a wide variety of familiar and unfamiliar topics. Other aspects to take into consideration when to assessing the candidates' vocabulary level include evaluating their ability to paraphrase successfully, usage of idiomatic vocabulary and sensitivity to register.

Fluency: The ability to speak English at an appropriate tempo, with only occasional loss of fluency is the main characteristic of a candidate who gets awarded a level 4 in fluency. A level 5 candidate is able to speak at length with relative ease, whereas a level 6 speaks at length with a natural, effortless flow. Other aspects to take into consideration when rating fluency are use of fillers, effectiveness of communication, use of discourse markers and speech variance for stylistic effect.

Comprehension: The comprehension descriptors talk about how accurate comprehension of common, concrete and work related topics is and about the candidates' ability to understand a linguistic or situational complication or an unexpected turn of events. The capacity of the candidates to understand a range of accents should also be evaluated. Level 6 descriptors include comprehension of linguistic and cultural subtleties.

Interactions: The main characteristic of evaluating interactions refers to the evaluation of the frequency in which candidates give immediate, appropriate and informative answers. The descriptors for interactions also include assessing how well candidates manage the communication (for example, if they deal adequately with unexpected situations such as misunderstandings). The level 6 descriptors also take into account sensitivity to verbal and non-verbal cues.

Appendix E: ICAO glossary of basic and complex structures

The structures compiled here are based on research at the Eurocontrol Institute of Air Navigation Services, Luxembourg.

Basic structures:

- Articles
- Adverbs of frequency
  - Always, Generally, Usually, Often, Sometimes, Seldom, Never, etc.
- Comparison of adjectives
- Discourse markers
  - Actually, Basically, Anyway, (and) yeah (more and more frequent), Listen, I mean, Let's see/Let me see, Like, Oh, Now, Okay, So, Well, You know, You see, You know what I mean, It is true, Of course, But, Still, (and) by the way, Besides, Another thing is, On top of that, So, Then, First(ly), Second(ly), etc., First of all, In the first/second place, Finally, In the end, In short
- Modal verbs
  - Can, May, Must, Have Got to, Should, Ought to, Would, Could, Might, Needn't, Don't have to, Mustn't
- Numbers (cardinal and ordinal)
- Passive voice
  - Simple present
  - Simple past
- Position of direct and indirect objects:
  - Bob sent some flowers to his girlfriend.
  - Bob sent his girlfriend some flowers.
- Question words for describing people and things and for requesting information
  - What? Who? Which? Why? Where? How?
- Relative pronouns
  - Who, which, whose
- Tenses
  - Present simple
    - I do
  - Present continuous
    - I am doing
  - Past simple
    - I did
  - Past continuous
    - I was doing
  - Present perfect simple
    - I have done
  - Present perfect continuous
    - I have been doing
  - Simple future tense
    - Will
  - Future
    - Going to
- There to be
  - Present, past, future

Complex structures

- Adjectives
    - Gradable and ungradable adjectives
        - Fairly angry (gradable)
        - Totally amazed (ungradable)
    - Prepositions after adjectives
        - Angry about, afraid of, etc.
    - Adjectives + that clause or to + infinitive
        - Enough, sufficiently, too + adjective
        - The sooner the better, etc.
- Adverbs and conjunctions
    - Comment adverbs
        - apparently, frankly, rightly
    - Viewpoint adverbs
        - biologically, ideologically, morally
    - Adverbial clauses of time
        - before, until, after, as soon as, before, when, while, hardly, no sooner, scarcely
    - Giving reasons
        - seeing that, since, in as much as, due to, owing to, with so many people ill
- Clauses
    - Relative clauses
    - Participle clauses
        - -ing, -ed and being -ed
    - Participle clauses with adverbial meaning
        - Opening her eyes, the baby began to cry.
        - Formed 25 years ago next month, the aviation club …
- Conditionals
    - Real and unreal, all tenses
- Discourse markers
    - Mind you, On the whole, Broadly speaking, By and large, Certainly, May, stressed "Do", On the one hand, On the other hand, While, Whereas, However, Even so, Nonetheless, Nevertheless, All the same, Although, Though, Even though, If, In spite of, Despite, Incidentally, Moreover, Furthermore, In addition, Additionally, (and) what is more, Therefore, As a result, Consequently, (Quite) on the contrary, To begin with, To start with, For one thing, For another thing, In conclusion, Briefly
- Infinitives and gerunds
- Modals
    - Will and would to show willingness, likelihood and certainty
    - Will and would to show habits
    - Modals + past participle to express criticism or regret
- Nouns
    - Compound nouns
    - Uncountable nouns with zero article
        - e.g. good advice
- Passive voice
    - Present perfect/past perfect/future/continuous forms in general

- Phrasal verbs:
  - They wanted to get the meeting over with.
  - The programme's lack of success could be put down to poor management.
  - Boeing came in for a lot of criticism over their new plan.
- Quantifiers
  - One of + plural
    - One of the best things
  - Each (of) and every + singular verb except when follows the noun or pronoun it refers to.
- Questions
  - Reporting questions
  - Negative questions
  - Question tags
- Reflexive pronouns
  - Herself, himself, themselves
  - One and ones
    - There’s my car — the green one.
  - So
    - I think so.
    - So I hear.
  - Do so
    - She won the competition in 1997 and seems likely to do so again.
  - Such
    - Such behaviour is unacceptable in most schools.
- Reported speech
  - They promised that they would help him the next day.
  - He told me it wasn't going to be ready by Friday.
- Verb tenses
  - Past Perfect
    - I had done
  - Past perfect continuous
    - I had been doing
  - Present continuous
    - For the future
  - Future continuous
    - I will be doing
  - Future perfect
    - I will have been doing
  - The future seen from the past
    - was going to, etc.

Appendix F: Interview guide

Interview guide (semi-structured interview)

Before the interview: check if everything is working well.

Interviewer: Good morning / good afternoon! First of all, I would like to thank you very much for agreeing on participating in this research. Your help is very much appreciated! The aim of the study is to investigate English expert and subject matter expert raters and test developers' experiences and opinions about the ICAO language proficiency requirements, its rating scale and the explanation of the descriptors. As you know, all the information collected about you during the course of the research will be kept strictly confidential. All collected data will be anonymised and your identity will not be revealed at any point.

Remember :
- pilot the interview questions;
- listen carefully to what the interviewee is saying;
- make probing questions;
- make supportive noises (back-channeling) ;
- speak as little as possible;
- encourage as much elaboration as possible from the interviewee

MAKE SURE THE INTERVIEWEE HAS EASY ACCESS TO THE SECOND EDITION OF THE DOC 9835 DURING THE INTERVIEW.

Interviewer: Before we start, I would like you to tell me a little bit about your professional background.

Background probing questions:

- Are you an English language expert (ELE)? A subject matter expert (SME)? Both?
- If SME: Pilot? Controller? Both? How many years of experience with international flights?
- If ELE: what is your academic background? Any experience with teaching English as a second language?
- Are you a rater? How long have you been working as a rater?
- Are you a test developer? How long have you been working as a test developer?
- Do you consider yourself to be an experienced rater/test developer?

Interviewer: OK, thank you very much. In the first part of this interview, I am going to ask you some questions related to the ICAO language proficiency requirements in general. Then, in the second part of the interview, I am going to ask you some specific questions about the ICAO rating scale. Do you have any questions so far?

Questions about the ICAO language proficiency requirements in general:

1) Do you think the target language use domain should only be the English used in communications between pilots and air traffic controllers? Why?/Why not?
2) Do you agree that the tests should be designed to assess speaking and listening? Why?/Why not?

3) Do you agree that the purpose of the test should be to assess plain language proficiency in an operational aviation context? Why?/Why not?
4) Is English phraseology tested in your country? If yes: do you know how it is tested? If no: do you think it should be formally tested? Why?/Why not?
5) Do you agree that phraseology should be tested separately from plain language? Why?/Why not?
6) Do you agree that responses containing elements of ICAO phraseology should not be rated with regard to their procedural appropriateness or technical correctness during the test? Why?/Why not?
7) Do you agree that technical knowledge of operations should not be evaluated during the test? Why?/ Why not?
8) Do you think operational level 4 is enough for aviation safety? Why?/Why not?
9) Do you think that it is adequate to retest candidates who got level 4 every 3 years? Why/Why not?
10) Do you think it is appropriate to retest candidates who got level 5 every 6 years? Why?/Why not?
11) Do you agree that level 6 pilots and air traffic controllers don’t ever need to be tested again? Why?/Why not?
12) Do you agree that a test –taker who is tentatively considered to be a level 6 speaker of the language may be evaluated through informal assessment (for example, by a flight examiner or licensing authority)?
13) Do you agree that the six categories that should be tested are: pronunciation, structure, vocabulary, fluency, comprehension and interactions? Is there any category that should not be in the rating scale? Is there any other category that you would include in the rating scale?
14) Which are the most important categories in your opinion? And the least important? How would you rate them from the most important to the least important?
15) Do you agree that the candidate’s final level should be the lowest level in any of the categories? Why?/Why not?

Interviewer: We are now going to talk about the rating scale. Before we go on with the interview, would you like to say anything else about the language proficiency requirements in general?

16) What do you think are the strengths and weaknesses of the descriptors for pronunciation?
17) What is your opinion about the explanation of the pronunciation descriptors?
18) What do you think are the strengths and weaknesses of the descriptors for structure?
19) What is your opinion about the explanation of the structure descriptors?
20) Are you aware of the glossary of basic and complex structures published by ICAO in the second edition of the Document 9835? If aware, what is your opinion about it?
21) What do you think are the strengths and weaknesses of the descriptors for vocabulary?
22) What is your opinion about the explanation of the vocabulary descriptors?
23) What do you think are the strengths and weaknesses of the descriptors for fluency?
24) What is your opinion about the explanation of the fluency descriptors?
25) What do you think are the strengths and weaknesses of the descriptors for comprehension?
26) What is your opinion about the explanation of the comprehension descriptors?
27) What do you think are the strengths and weaknesses of the descriptors for interactions?
28) What is your opinion about the explanation of the interactions descriptors?

29) Is there anything that you think should not be in the rating scale descriptors? If so, what? Why?
30) Is there anything that you think should be rephrased/modified? If so, what? Why?
31) Is there anything that you think is missing in the rating scale descriptors? If so, what? Why?

Final questions:

What should I have asked you that I didn’t ask?

Is there anything more that you would like to bring up, or ask about, before we finish the interview?

Interviewer: This is the end of the interview. Thank you very much!

Appendix G: Interview sample

Interviewer: Good morning, here it's good morning. It's good afternoon for you!

Participant: Indeed, indeed, but only just.

Interviewer: First of all, I would like to thank you very much for agreeing on participating in this research. Your help is very much appreciated! The aim of the study is to investigate English expert and subject matter expert raters and test developers' experiences and opinions about the ICAO language proficiency requirements, its rating scale and the explanation of the descriptors. As you know, all the information collected about you. All the collected information and data will be anonymised and your identity will not be revealed at any point. The information will be kept strictly confidential.

Participant: Ok, lovely, that's fine, that's absolutely fine. Of course if you want to say what I've said you're very welcome to, it's no problem.

Interviewer: All right. Do you have the doc with you, the rating scale and the explanation of the rating scale descriptors on the doc, or you don't have that?

Participant: Let me open it up now. So we're looking into document 9835, right?

Interviewer: Yes, chapter 4, section4.6. But just keep it close to you, first we're going to start talking about the language proficiency requirements in general, then we're going to talk about the scale and the explanation of the descriptors.

Participant: OK, that sounds good.

Interviewer: Before we start with the questions, I would like you to tell me a little bit about your professional background, please.

Participant: OK, I've been working in this area, aviation English since about 2002. I've been in language teaching and assessment all of my professional life, I haven't done anything else. I started in 1980 with a certificate in English language teaching and then got a DELTA in 2005. The two books that you probably know, published in 2008/2010 led the ICAO rated speech samples training aid I was training aid which was published in 2012, developed the English test for , are we being recorded now?

Interviewer: Yes, yes.

Participant: Good so you've got this on as a voice phone. I was just aware I was going quite quickly then. And led to the development of the English test for aviation, which was the first test to receive recognition, or at the time it was called endorsement, from ICAO. I guess the aviation English has been central to my professional life since I started working in this area in about 2002

Interviewer: OK, thank you very much. In the first part of this interview, I am going to ask you some questions related to the ICAO language proficiency requirements in general. Then, in the second part of the interview, I am going to ask you some specific questions about the ICAO rating scale. Do you have any questions so far?

Participant: No, I don't think so.

Interviewer: Alright, so let's start. Do you think the target language use domain should only be the English used in communication between pilots and air traffic controllers or should it include other kinds of communication, you know, between pilots and other pilots, pilots and flight attendants, pilots and mechanics or controllers and other controllers, supervisors. What do you think?

Participant: I think it is probably best to restrict it to air/ground communications and not extending it to other target language use domains, for example, cockpit/cabin or pilot to engineer or pilot to dispatcher or the pilot to passenger. They are all so varied and so different, that it would be difficult to capture the essence of all of those multiple uses of language and all of the variation in one set of criteria for assessment. Keep it restricted and in my view that will make things easier for test developers and for assessors to deal with.

Interviewer: Do you agree that the tests should be designed to assess only speaking and listening? Why? Why not?

Participant: I think speaking and listening is fairly reasonable to test takers, as that is the nature of air/ground communications and, again, if it were extended to other skills, such as reading and writing, you would be taking things far beyond the context of air/ground communications and that would make things significantly more complicated.

Interviewer: Do you agree that the purpose of the test should be to assess plain language proficiency in an operational aviation context?

Participant: No, I don’t agree with that. I believe that the intention of ICAO is to assess language proficiency in the context of radio communications and that target language use domain is made up of two really important components, first being standard radiotelephony phraseology and the second being plain English where phraseology doesn’t suffice. So I think that ICAO missed an opportunity to combine those two elements, the target language use domain and I think, had they done that, had they encouraged test developers to work on both phraseology and plain language together, then on one hand we would have stronger tests because naturally you would want to engage the ability of the candidate to improve and use, demonstrate their ability to use both standard radiotelephony phraseology and plain English so the quality of test would be better and would be linked more to air/ground communications and I think secondly, it would have helped ICAO to address the issue of proficient uses of English insofar as I quite strongly believe that a lot of communication problems are to do with poor use of phraseology amongst native speaking crew and possibly air traffic control too, that’s more I think based on anecdotal evidence than any hard evidence from research into the area but it would seem that standard phraseology, if it is used correctly, would contribute significantly to safety in communications. So I think it should include both phraseology and plain English, to separate the two is to artificially divide a single construct which is safe pilot/controller communications into constituent parts which don’t necessarily want to be divided. Sometimes it is very difficult, for example, to see where phraseology ends and where plain English begins. The switch between the two happens so fluidly among proficient users. And I think the distinction between the two is sometimes very, very difficult to see.

Interviewer: And Is English phraseology tested in your country? Do you know?

Participant: Yes, it is. We have the UKCAA flight radio telephony operators’ license. That is a test which pilot trainees take at a late stage of their training in order to gain the license to use the radiotelephony and that’s separate from other aspects of personal licensing. So, yes, phraseology is tested.

Interviewer: Do you agree that responses containing elements of ICAO phraseology should not be rated with regard to their procedural appropriateness or technical correctness during the test?

Participant: I disagree with that again. I think the ability to communicate effectively in an radiotelephony context is dependent on A) accurate use and appropriate use of standard radiotelephony phraseology and B) where that phraseology doesn't suffice, good command of concise, brief and clear plain English. So I think being a proficient user of the radio requires skills on both sides and I think tests should be measuring both of those things in tandem together because they are part and parcel of the same construct. That's my opinion I appreciate that ICAO guidance has us test plain English separately, but I don't think that is the right way of dealing with radiotelephony communications.

Interviewer: What about technical knowledge? Do you agree that technical knowledge of operations should not be evaluated during the test?

Participant: I think it's impossible not to evaluate technical knowledge during a test, for example, if you have in a well-developed task which simulates R radiotelephony communications, if you give a pilot or a controller a scenario which engages plain language use in that context, you cannot separate procedural or operational knowledge from that language performance, for example, if a pilot is talking about hydraulic loss, engine problems or weather issues, the way that he or she chooses vocabulary, chooses how to wrap up their meaning into plain English communications, will be dictated by their knowledge of operational procedure. Talking about, for example, hydraulic loss on a radio is very, very different from talking about hydraulic loss in a broader sense, in a test room or with another pilot, if you are having a beer or whatever. It's very specific and it's very linked to the operational context in which the traffic is operated. So, I think it's impossible and undesirable to try and separate operational knowledge from language knowledge.

Interviewer: And do you think operational level 4 is enough for aviation safety?

Participant: Really good question and the answer is I don't know and I would be reluctant to volunteer an opinion without seeing that opinion supported by some solid research in the area. My gut feeling is yes, it would do as a minimum level, but I think we really need to understand today how pilots and controllers who today by and large have an ICAO level 4 plus perceive the effectiveness of communication. For example today if you ask pilots or controllers whether the people they talk to on the radio are good enough to do the job, this will tell us whether ICAO level 4 is functioning, as intended in terms of establishing a minimum level of proficiency for safe communications.

Interviewer: That’s a good suggestion for further research

Participant: I think it would be really good research to look into actual perceptions of those using the language whether everybody has achieved the right level.

Interviewer: According to your experience, the pilots you know and you rated do you think that it is adequate to retest candidates who got level 4 every 3 years?

Participant: That’s another good question. The policy is 3 years, how closely the policy aligns with actual language decay, I don’t know. It would seem reasonable every 3 years to spend 30 minutes, 45 minutes whatever doing a language test is a mere drop in the ocean in the amount of time pilots and controllers spend in ongoing training if you have pilots doing operational proficiency checks every 6 months why not incorporate language proficiency into that operational proficiency check, for example? But 3 years would seem reasonable. Again, I think it needs to be supported by evidence from non-test real life language use to see whether it’s enough or not.

Interviewer: What about candidates who got level 5 being tested every 6 years?

Participant: Again I think the same comment would apply, I don’t know. It seems to be reasonable, but would need to be borne out by evidence.

Interviewer: Do you agree that level 6 pilots and air traffic controllers don’t ever need to be tested again?

Participant: No, I absolutely disagree with that. Being a highly proficient user of English doesn’t mean you are a highly effective user of the RT, particularly when you have students coming through an English medium aviation training program, let’s say, for example, Brazilian students learning to fly in Florida or Chinese students learning to fly in the United Kingdom. They will acquire quite readily a very high level of language proficiency on account of being immersed in an English speaking environment during training. That doesn’t mean that being a level 6 on completion of training will mean that that student is level 6 in 15 years’ time, depending on where he or she operates. But I think it’s unfair for proficient users of English to be exempt from an ongoing license requirement, for example, if we look at aviation medicine, if you pass your medical extremely highly, you’re flawless in terms of your physiology, it doesn’t mean you’re going to be flawless in 10 years’ time, so you would never be exempted for life on medical grounds, why be exempted for life on the grounds of effective communication. I also think that there is a political dimension to this, and that native speakers are often perceived to be level 6, but being a native speaker certainly doesn’t mean you are an effective user of English as lingua franca in the aviation context.

Interviewer: That relates to the next question. Do you agree that a test –taker who is tentatively considered to be a level 6 speaker of the language may be evaluated through informal assessment, for example, by a flight examiner or licensing authority?

Participant: I can see the enormous attraction of doing that, but, again, language tests are developed with a specific purpose in mind and that in our context is to make valid influences about the ability of a pilot or a controller to communicate effectively. So, I think that an informal testing context will not engage the abilities that we are looking for in safe radiotelephony communications, particularly that very important switch between standard phraseology and plain English. If, for example, you are in a simulator, you do a series of simulator exercises and then you have a debrief with your simulator instructor in order to be signed off and your license revalidated, the discussion you have with that instructor is not going to engage the range of listening comprehension abilities or the ability to switch between standard phraseology and plain English, so, no, I strongly disagree that informal testing of high level users of English is an acceptable means of determining the ability to communicate safely on the radio.

Interviewer: And how do you understand the holistic descriptor that says – I'm sorry this is not a question in my interview guide, but something that came up to my mind - the holistic descriptor says that 'test takers should demonstrate ability to communicate in face-to-face situations', but the TLU is the radiotelephony communications, there is no face-to-face communications in that. Why do you think they did that? Do you think there is a reason for that face-to-face communications, although the TLU doesn't have face-to-face communication?

Participant: No, I have no idea, no idea at all. You could, for example, argue that a controller communicating with a pilot and taking his or her information and then passing it on to the next sector might require the controller to stand up and walk to the other side of the control room and say watch out for this guy he's got an issue with a passenger or whatever, but that is not part of the target language use domain and I think the way that you phrased your question there was 'do have any idea why they did that'. I would suspect that you would agree with me.

Interviewer: Yes, I don't understand that.

Participant: No, neither do I.

Interviewer: OK, so we are now going to talk about the rating scale. So, before we go on with the interview, would you like to say anything else about the language proficiency requirements in general?

Participant: No, I think other than issues like the one you just raised, the fact that face-to-face communication is mentioned in the holistic descriptors, I think this only leads to confusion over what we are testing and how we are testing it, so I think likewise ICAO's guidance that it's acceptable to test high level users in an informal setting, I think this, again, dilutes the message that we are testing English for a very, very specific purpose, for safe communications. I think it confuses test developers and that confuses authorities and I think it also confuses test takers. It's not uncommon for pilots to say why are we doing this? I never do this as part of my job. So, I think altogether the policy doesn't fit particularly well the target language use domain, but, that's just my opinion.

Interviewer: Ok, thank you! Do you agree that the six categories that should be tested are: pronunciation, structure, vocabulary, fluency, comprehension and interactions or is there any category that should not be in the rating scale, or is there any category that you would include in the rating scale?

Participant: I don’t think there are any other categories that we should include in the rating scale, but I don’t believe comprehension should be in the rating scale.

Interviewer: Why not?

Participant: Because comprehension sitting alongside components of spoken language proficiency firstly diminishes the importance of listening comprehension. If we consider listening/speaking to be skills which are equal and they interact and relate very closely together the way that comprehension is perceived in the rating scale is perceived as one of six things that students should be able to do, when it is not, it’s one of two things that students should be able to do, or pilots and controllers should be able to, one being speaking and the second being comprehension. So I think it misleads us to think that comprehension is a very thin slice of the ability to speak and it is not, it’s an extremely important, if not more important, part of the overall proficiency construct in this case.

Interviewer: So what would you suggest? How should it be?

Participant: I think comprehension could be its own set of scales. You could expand comprehension criteria to cover all of the things you expect pilots and controllers to be able to understand and you could expand that into a lot of detail, it might be understanding different accents, it might be understanding different scenarios, it might be understanding, for example, the difference between alphanumeric information as is contained in call signs or flight levels or headings, whatever, as different from phraseology, as different from the sort of information you would see in plain English in non-routine situations. And I don’t think the listening construct in the case of radiotelephony communications is adequately defined or captured by the criteria as they stand.

Interviewer: That’s interesting.

Participant: I also think that it leads test developers to assume that it is possible to test listening comprehension in an oral language test, so sitting down and talking with somebody about what they do and maybe engaging in simulated pilot/controller communication or whatever the task might be, is never going to tap the listening construct fully as is necessary to distinguish between levels 4, 5 and 6. If for example you get up to level 6, and you’ve got the ability to understand cultural subtleties, how are you going to get comprehensible input textual material looking to an oral proficiency test which contains a range of cultural subtleties to allow you to distinguish between level 5 and level 6. It requires a listening test which is separate from the candidate’s ability to speak. So I think comprehension is particularly poorly dealt with in the criteria.

Interviewer: Would you be able to rate the categories from the most important to the least important?

Participant: Yeah, I would put listening comprehension as level 1, because without your ability to listen and understand what is happening you have no chance of using your spoken language performance to engage in communication. I would suggest that pronunciation plays a very high part as well, alongside interactions and I would suggest that structure plays much less of a role, or at least the structure as it is captured by the rating scale. I would say fluency plays less of a role as well. Vocabulary is also important. This is the first time I’ve ever thought about that, Angela, I don’t have a good answer for you, but I would say maybe comprehension, pronunciation and interactions working together as a very close second, and then, in third place, vocabulary, fourth place fluency, fifth place structure.

Interviewer: Thank you. And do you agree that the candidate’s final level should be the lowest level in any of the categories?

Participant: Yeah, I think that’s reasonable. I think there are two good things about that. Firstly, you take the lowest common denominator, so the chain being strong as the weakest link, that old cliché, I think it’s true, but also I think it helps to improve test reliability in that two examiners are much more likely to agree on the overall operational proficiency of a candidate when you take into account the lowest of any of the scores in each of the criteria. So I think it improves test reliability, but it also improves safety.

Interviewer: All right, thank you, now let's talk about the descriptors. The first question is going to be about the descriptors and then the explanation of the descriptors. Do you have the manual opened on chapter 4.6?

Participant: I do.

Interviewer: OK. So what do you think are the strengths and weaknesses of the descriptors for pronunciation?

Participant: I think an enormous weakness is reference to the candidate's first language. It's a bit unfortunate that the candidate's first language is considered to be a negative, even at the very highest levels of language proficiency candidates very, very often display features of their first language pronunciation and I certainly don't see how this should be considered to be a problem. There's nothing wrong with it. I think it also gives candidates a sense that they shouldn't be speaking in English with any influence of their first language. I think that's very unfortunate and it's quite culturally insensitive too. You take some level 6 users of English, they are extremely proficient, and they are very obviously not native like speakers, they have influence. So I think that's a negative. Another negative is separating pronunciation in terms of frequency that pronunciation interferes with ease of understanding. Really the only thing that separates 3 from 4, 4 from 5, and so on, is the difference in words like 'sometimes', 'usually', 'rarely'… And this is very difficult to quantify, it remains subjective. However, I do think that the phrasing of ease of understanding is very positive. I think that's good. It's the understanding the impact that it has on the listener, I think is a good way of looking into how effective pronunciation is, so I think that is a positive.

Interviewer: What is your opinion about the explanation of the descriptors?

Participant: I'm just reading them here, so bear with me.

Interviewer: Take your time. When I ask that, I'm asking what helped you understand the descriptors, what made you feel more confused about descriptors or if there is something that contradicts the descriptors.

Participant: You see, I am just feeling irritated by this explanation. It says 'it should be noted that native or second language speakers may be assessed at this level', this level being level 3, 'in cases where a regional variety of the language has not been sufficiently attenuate'. Now my English is, the English that I personally speak, is very standard British English and anybody talking to me would say, 'you're from Britain and you're from the South of England'. Now, I've never made any attempt to attenuate the language that I use, why should we? And I find it irritating that there is a perception that there is a model of English which isn't linked to a particular part of the world. It is simply untrue, wherever you are from, you speak English with an accent. There is no such thing as a neutral English accent. It's only either British or American or Australian or Brazilian or Chinese or whatever it might be. And I find it irritating that this idea of accent reduction somehow helps pronunciation to be more effective. For example, Angela, you speak brilliant English, your pronunciation is wonderful, I never have any issue understanding you at all, your pronunciation is simply fantastic. But you are a Brazilian Portuguese speaker of English, that is very, very obvious, and I would never want to say to you or to any other speakers of English: lose your accent, because A) why would you want to do that? It's your cultural heritage where you're from, it's your identity, but secondly it doesn't impact on my ability to understand you. So I think it's a nonsense to introduce this, sorry I find it irritating. I think it's really unfortunate.

Interviewer: I agree. Well, is there anything else related to pronunciation that you have experienced to be difficult in test development or in rating?

Participant: No, I don't think so, in term of test development or in rating. In rating, yes, because the jump from frequently interfere with ease of understanding to sometimes interfere with ease of understanding, level 3 and level 4 is an enormous jump. You can't be a very strong 3 and a very weak 4 and be separated by descriptors which are so widely different.

Interviewer: Now what about structure? What do you think are the strengths and weaknesses of the descriptors for structure?

Participant: I like the way the focus is on global and local error. I like the way that level 6 allows the user to make errors, that is consistently well-controlled doesn't mean error free, and I think that is very good, because as highly proficient users of the language we make errors with it all the time. So I think that's good. I think that there are issues with the jump from level 4 to level 5 and those would be at level 5 complex structures appear. Does that mean that level 4 users never attempt complex structures? I think that there is a very delicate and poorly understood interplay between basic and complex structures, for example at level 5, it says basic structures need to be consistently well controlled with complex structures with error, it's very rare to come across candidates who do that. It's much more frequent to come across candidates who consistently make few errors with their basic structures and attempt complex structures and sometimes they get that right as well, but it's still got error in it. So I think this relationship between basic and complex structures is poorly understood and from what I've read in the literature, it is actually extremely difficult to categorize structures as basic or complex particularly when you're dealing with international uses of English where for some speakers a particular structure in English may be highly complex because there is no equivalent in their first language and the meaning of that structure is not expressed in their first language, so using it accurately and appropriately is extremely difficult. However, that structure may appear in ICAO's list of simple structures and a good example here is the present perfect, for some speakers of English the present perfect is a nightmare and is always a nightmare, it doesn't matter how good you are in English present perfect is really difficult to get right and yet ICAO listed as a simple structure. So I think that simple versus complex dichotomy is poorly understood and poorly worded.

Interviewer: And would you like to make any comment on the explanation of the descriptors or other about the glossary of basic and complex structures published by ICAO?

Participant: This is interesting in the explanation that says 'level 4 speakers will not usually attempt complex structures and when they do quite a lot of errors will be expected resulting in less effective communication'. That is not written in the criteria.

Interviewer: So do you think that helps?

Participant: Yeah, it would help, if that was actually written, something like, a level 4 user may attempt complex structures, but they will have errors which interfere with meaning, but if you write that at level 4 you are actually duplicating the level 5 descriptor. So, yeah, I think sometimes. To be honest, Angela, I think the rating scale was very poorly thought out. This is not a criticism of ICAO, they had to do a job and they did, and they called in the experts they had available, and they had very limited time and a tiny budget to get this right and they didn't get it right. But I would be a strong proponent of rating scale revision based on actual language use rather than prescriptive language use and I think what ICAO has done with these explanations is to add a layer of complication rather than clarifying, they actually complicate and confuse. That's again my opinion.

Interviewer: So now let's talk about vocabulary. What are the strengths and weaknesses of the descriptors? Any comments on the explanation of the rating scale? In terms of rating and test development?

Participant: Just to go back to structure just for a second. The list of basic and complex structures that ICAO has drawn up, that's refer to their Appendix B Part IV, it is not rooted in any sense of research either in the target language use domain or in the wider field of applied linguistics language teaching, and I think it's really poorly thought out, for example, they have in their aspects of vocabulary, for example being able to grade adjectives to say it's fairly hot, but you can't say it's absolutely hot, that's an issue with vocabulary, that's not grammar, at all, and it's the grammar of vocabulary, it's lexical grammar, whatever you want to call it, but it's just… It's really, really poor, anyway I think to be irritating to be working with something which is, you know, just badly thought out. Vocabulary I think the glaring issue with that and I'm sure you'd agree with me it is the appearance of idiomatic language, number 5. It's got no place in radiotelephony communications. It doesn't necessarily identify strong users from weaker users. It has a deleterious effect on safety and it shouldn't be there, it has absolutely no place in this rating scale

Interviewer: I totally agree with you.

Participant: I thought you might. And then at level 6 we've got this 'vocabulary is sensitive to register'. What does that mean? We are talking about one register and that's the ability to communicate on the radio. You don't have multiple registers on the radio. It's short, brief, concise, to the point, safety operational related language use. There is no room for different registers in that context, so it is nonsense to include it in the scale. And I think for test developers it makes it very difficult because it means to engage this level 6 ability properly in a language test you have to do things which are irrelevant to the target language use domain and I think this dilutes the message to the personnel, the people who have to take the test, and I think it is a threat to the quality of language testing.

Interviewer: Any other comments?

Participant: No, I don’t think so.

Interviewer: Do you see a difference between the common, concrete and work related topics, are they referring to the same kind of topics or is there a difference between them?

Participant: Yeah, good question. I think this comes across in the scale in other criteria as well. I think common, concrete and work related topics are all different things. If they are meant to be the same thing, why not use one word? If something is common, it happens very frequently and if something happens very frequently, it’s usually very safe, so, for example, an aircraft taxies to the holding position. That happens commonly and there are checklists and there are phraseologies for that particular maneuver. So why you would have language which is common, that is not related to that... I think again it’s very confusing. If something is concrete, you understand it, you lost a door and you lost pressure as a result or you’re having problems intercepting the localizer, or this particularly bad weather visibility is below minima on takeoff... Those are concrete situations. They might not be common, but they are concrete. And then we’ve got work related. Well that could be anything to do with, you know, from checking into a hotel once you arrive at your destination to programming the FMS at the beginning of a flight, if it’s work related it’s anything to do with the job. So I think they are all different things and I think they are poorly, again, poorly conceptualized.

Interviewer: Alright. So thank you very much, let’s move on to fluency. What are the strengths and weaknesses of the descriptors, in your opinion?

Participant: I think descriptors being able to speak at length, again, it never happens on the radio, why is it in there? I think at level 6 there are issues with the descriptors around being able to vary speech rate for stylistic effect, for example, to emphasize a point. The ability to control speech rate to emphasize a point is linked very much to the emphasis that you place on the word by stressing a particular syllable and adding increased phonetic value to that syllable and it’s much more close in link to features of pronunciation than it is to fluency. I think that is a really poorly conceptualized aspect of the fluency descriptor. Speaking at length with relative ease, well, ok, that is fine insofar as it distinguishes between level 4 and level 5, but pilots and controllers never do it on the radio, so what is it doing there? I think that at level 4 it is really good that there is this 'loss of fluency on transition from rehearsed or formulaic speech to spontaneous interaction'. That's a mouthful, what they mean there is there is a loss of fluency on switching from standard radiotelephony to plain English or plain language. If that is what they meant, why didn't they write that? It leads test developers to think well it is ok to test outside of the context of radiotelephony communications, which leads me to think how were they conceptualizing rehearsed or formulaic speech to spontaneous interaction? We have language routines, of course, 'how are you?', 'I'm fine', 'what about you?' 'yeah, I'm fine, thanks'. These are routines and we have routines all the time, but rehearsed or formulaic speech to spontaneous interaction, I think it is really poorly defined. It is quite interesting that in the fluency explanation, it doesn't refer to this and it keeps referring to 100 words per minute, 100 words per minute is so unnaturally slow. Pilots/controllers never do it, they never communicate 100 words per minute. They speak much more quickly than that. And I don't understand, yes, it is desirable, if eve-ry-bod-y-spoke-a-a-hun-dred-words-per-min-ute. It might be clearer, but it might also occupy an enormous amount of radio time, particularly when you are in busy airspace. Again, it is prescription rather than understanding what really happens, capturing that in the criteria.

Interviewer: Alright. Now the comprehension descriptors, weaknesses and strengths.

Participant: I think that, as I mentioned before, they are inadequately defined both in the criteria and in these explanations of the criteria. I think that they are overwordy, linguistic or situational complication. I think it is just, it is irritatingly wordy. If, for example, you've got a non-routine situation and the pilot calls up and he says "unable to maintain speed due weight, request lower speed, please". That is a situational complication, because there isn't a standard phrase to deal with it and that situational complication leads to a linguistic complication because you are moving outside phraseology and into the much more dangerous area of plain language use. So I think situational complications and linguistic complications co-occur, so they happen at the same time, one doesn't happen separately from the other. So I think they are over wordy. I think that the focus at level 4 and this is true right across the rating scale, on non-routine situations at level 4 is very good and level 4 really is I think the best of all of the levels as they are written in the scale in that it captures radio communications much better than others. There is nothing really irrelevant in level 4 right across the board to the target language use domain when you move up, possibly when you move down, it is a bit more irrelevant. I don't think I've really got anything more to add there. Non-standard dialects, or regional accents, again it is very difficult to conceptualize that in a language test.

Interviewer: I also think it is difficult.

Participant: Yeah, how much listening comprehension is enough to reliably distinguish between 3, 4, 5 and 6? If you take language proficiency tests in other contexts, if you want to distinguish reliably between all of those levels you are looking at a listening test which could last 45 minutes or an hour, so you've got enough items in there, all sufficient levels of difficulty to be able to distinguish with reliability between those levels. So again I think that the rating scale gives us the impression that comprehension is I guess, how many levels? There are 6 criteria in the scale, aren't there? Comprehension is one, so it makes us think that comprehension is less than 20% of the overall ability to communicate on the radio. It is not. It is 50%, at least. Do you see what I mean with that?

Interviewer: Yes, I agree. What about this part when they, pre-operational level 3, they say may fail to understand the linguistic or situational complication, I'm not talking about the linguistic or situational complication, may fail, is it clear, I don't know how many items there are in the test, but if the pilot or controller fails to understand one or two items, would he be a level 3 or a level 4, because a level 4, ok, is mostly accurate, but the pre-operational level 3 says may fail to understand and the candidate failed in some items, is it confusing for you when you are rating or developing a rating manual for the raters.

Participant: I think so. I think so because I think this is better measured in a separate test of listening comprehension in any case. Insofar as you can, for example, how difficult are linguistic situational complications? You might have a linguistic, situational complication that is really easy to understand, set a latitude 243, having a problem with a passenger, requesting immediate return to your airfield, that is easy to understand, problem, passenger, return. You might have a far more complicated situation to understand whereby, I don't know, you've got multiple system failures and the crew itself isn’t particularly language proficient, so not communicating the nature of their problem particularly clearly, which would lead to clarification and confirmation strategies. So a linguistic situational complication isn’t a single event at a defined level of difficulty. So this is why I think we need to have quite comprehensive tests with levels of difficulty of items within those tests, so we can make distinctions between what we think are easy linguistic situational complications and more difficult ones to help us distinguish between levels 3, 4 and 5 and 6.

Interviewer: When they say on level 4 that ‘when a level 4 is confronted with a linguistic situational complication the comprehension may be slower or require clarification strategies’, do you think that means they will in the end understand it after asking for clarification or even though they are a bit slower they will understand and a level 3 may fail to understand, so would that be one of the ways to distinguish a level 3 from a level 4?

Participant: If we listen to the air traffic control recordings at the Hudson double bird strike where Sully Sullenberger calls up and says in response to the controllers instruction ‘unable, we’ll be in the Hudson’. That controller comes back and he says his instruction again and Sully says again ‘unable, we’ll be in the Hudson’ and then the controller calls up again and he says this airfield is available for you if you want to and Sully says again ‘unable’. Now did that controller misunderstand? He is a native speaker, possibly he did. Possibly it was such a difficult situational complication to understand a double bird strike with the captain of the aircraft willingly saying he is going to land the aircraft in the water. That is difficult to understand, not because of language, but because of the situation. But of course it leads to linguistic complications because Sully is just saying ‘unable’ and the controller is thinking ‘what does that mean?’ So I don’t think we can determine the ability to understand in live interactions because what you say is a result of what you hear, doesn’t reflect what you’ve heard, it reflects what you think you need to say in the situation. So this idea that checking, confirming and clarifying is a feature that only lower level proficiency users display is wrong, because much higher level listeners also display that feature when there is a situation which is very difficult to comprehend.

Interviewer: And what is your opinion about the comprehension of cultural subtleties?

Participant: I think it is a very valid part of the construct because you do get a lot of language which is linked to culture, the two are inseparable, and you hear a lot of that in radio communications. What aspects of it to include in the test becomes extremely complicated to operationalize. What cultural subtleties do you include? Whose? It's really difficult that one… It is a slippery beast.

Interviewer: All right. Now what about interactions?

Participant: I think a lot of this ties in with comprehension. I think it is very unfortunate that this checking confirming and clarifying seems to be a feature of comprehension and a feature of interactions, you are almost weighing the same thing, that is the same part of the construct appearing in two different aspects of the rating scale. I think that is confusing. I think the upper levels are poorly defined. ICAO says a level 5 should be able to do everything that a level 4 does, plus a bit more and a level 6 can do everything that a level 5 does, plus a bit more, which is why maybe in interactions you've got such a thin description at the top level. I forgot exactly what it says – 'interacts with ease in nearly all situations'. That descriptor seems weaker than 'responses are immediate appropriate and informative'. I would rather have somebody that responds immediately appropriately and informatively talking to me than somebody who interacts with ease in nearly all situations, which implies that there are situations where they don't interact with ease at all. So I think this issue of verbal and non-verbal cues is complete nonsense. We know we don't have any benefit of eye contact in voice only communications so what that is doing in there I really don't know. To me a verbal cue is anything that comes from the voice. So if we are saying eee, ahumm, shuuuuuu, whatever sound we make that is a verbal cue. So it means that non-verbal cues are anything that we do to communicate not using the voice, i.e. gesture and body language. Firstly, gesture and body language, yes they are very important for communication and good communicators quite often gesture very well and use body language very well, but we are on the radio for goodness sake!

Interviewer: When they explain that in the explanation they say 'they are additionally able to recognize and to use non-verbal signs of mental and emotional states, for example, intonations or unusual stressed patterns'. Do you think that is verbal?

Participant: Yeah, of course it is it. Intonation is of the voice and unusual stress patterns is of the voice it is verbal. So what does non-verbal mean then?

Interviewer: What about this ‘deals adequately with misunderstandings’, they check, confirm and clarify’, level 4, and level 3, ‘generally inadequate when dealing with an unexpected turn of events’. Do you think it is hard to rate that or develop a test to check the misunderstandings, for example, in our test there are three misunderstandings in part 2, well, there are two misunderstandings and one confirmation and we are always in doubt when to give a level 3 or a level 4, we standardized that if the candidate can clarify two misunderstandings or one misunderstanding and one confirmation out of three, two out of three, he may be a level 4. But some people disagree, some people think that they should be able to confirm or clarify all the three. So what is your opinion about that?

Participant: It is a good question. It is really good to hear that you have done some work with standardizing your approach to that because it could be very subjective unless there were standards in place. I think it needs the decision of the test developer. You create standards and then you stick to it. It is difficult to say what the right way of approaching that would be, but so long as there is consistency in your approach, that is an important part of it. But I think it is a good descriptor, deals adequately and is inadequate. I think again it comes back in just a comment about the overall scale, I think the distinction between level 3 and level 4 is very clear, quite often the two levels are very far apart in the way they are described, but level 4 really does address much more closely the TLU, so I think that is a good feature of the way the scale is put together. I think it is unfortunate that there is a lot of repetition of terminology, for example, you’ve got predictable situations, and common, concrete, work-related situations, you’ve got situational linguistic turns of events, unexpected turn of events… Just choose one descriptor to capture what you actually mean, rather than using a variety of ways of describing that same thing, or at least to me it seems like the same thing. I think it depends really, coming back to your question, Angela, I think it depends how difficult those situations that you are asking the pilots to confirm are. And it is also quite difficult in a language test to introduce mock misunderstanding. Is the interlocutor in order to engage this checking, confirming and clarifying in live interactions, you have to pretend that you haven’t understood something that the candidate has said, and you make it clear that you haven’t understood without saying I didn’t understand that, so, for example, you give me a squawk and say “latitude 123”, “squawk 361”, and I say to you “squawking 316”. It is clear to you that I’ve got my squawk wrong, it is up to you to put it right and I would expect you as a test candidate to say, ‘negative, say again squawk

361’. In test development you have to try to introduce this sort of deliberate misunderstanding to see if the candidate can put it right, and that is difficult.

Interviewer: So if you were to revise the rating scale, you mentioned you would write the descriptors in terms of language use. You’d change like almost everything.

Participant: Yeah, I think I would.

Interviewer: You agree that the scale needs to be revised.

Participant: Absolutely!

Interviewer: OK, we are getting to the end of the interview now. What should I have asked you that I didn’t ask?

Participant: I don’t think there is anything, Angela. You’ve given a really comprehensive interview.

Interviewer: Is there anything more that you would like to bring up or ask about before we finish the interview?

Participant: I think it would be really useful to have some suggestions as to how the scale might be revised, what procedures, what methodologies would be in place who would help do that and what is the safety case for that. ICAO is always saying ‘we can’t do this and this, that is a safety case’. Having an unreliable scale is enough of a safety case for them. I think that the research you do if you can share it as widely as possible with the aviation English community, circulate it, distribute it , publish it, we’ll be very happy somehow via our channels to say have a look at this document it is really good, everyone is going to read it because the work you do is important.

Interviewer: Oh, thank you. I hope I can help.

Appendix J: Other relevant strengths and weaknesses of the descriptors

Pronunciation

Strengths	Weaknesses
"Accent at this pre-operational level 3 is so strong as to render comprehension by an international community of aeronautical radiotelephony users very difficult or impossible" (explanation for level 3), mentioned by participants B and D. Participant B mentions it helps because it helps to differentiate a level 3 from a level 4, not only depending on the adverbs of frequency.	"An Expert Level 6 speaker may be a speaker of English as a first language with a widely understood dialect or may be a very proficient second-language speaker, again with a widely used or understood accent and/or dialect" (explanation for level 6). Participant E questioned: "is there an assumption that if something is widely used, it can be understood?"
"Only occasionally a proficient listener may have to pay close attention" (explanation for level 5), mentioned by participants B and E.	The category is "Pronunciation" and the descriptors talk about "pronunciation, stress, rhythm and intonation", but it is not clear what "pronunciation" means, as mentioned by participant B. The term pronunciation should be replaced by a clearer terminology, for example, "individual sound segments"
"Expert speakers are always clear and understandable" (explanation for level 6), mentioned by participant E.	

Structure

Strengths	Weaknesses
Level 6 descriptors allowing for errors, as mentioned by participant C. He argued that “consistently well-controlled doesn’t mean error free, and I think that is very good, because as highly proficient users of the language we make errors with it all the time”.	Language functions are not in the descriptors or explanation, only in the introductory paragraph for structure.
“Level 4 speakers will not usually attempt complex structures and when they do quite a lot of errors will be expected resulting in less effective communication” (explanation for level 4), mentioned by participants B and C. It helps because the descriptors for level 4 only talk about basic structures, and there are candidates at this level who produce some complex structures. However, it should have been included in the descriptors.	The adverbs of frequency in the descriptors may lead to different interpretations, mentioned by participants B and D.

Vocabulary

Strengths	Weaknesses
"Gaps in vocabulary knowledge and/or choice of wrong or non-existent words are apparent at this level. This has a negative impact on fluency" (explanation for level 3), mentioned by participants B and D. It helps because it complements the descriptors.	The difference between common, concrete and work-related topic is not clear, mentioned by participants B and C.
	The criteria do not explain how to deal with code-switching, mentioned by participant E She said that "I've had one or two switch to French, which is a strategy if you know the other person knows your language. But in an English test, I just assume that to be an error, but…"
	Not clear how to rate a candidate who does not need to paraphrase during the test, mentioned by participant D.
	The reference to unfamiliar topics on level 6 goes beyond the TLU domain, mentioned by participant B

Fluency

Strengths	Weaknesses
Reference to effective communication, mentioned by participant B.*	"Can make limited use of discourse markers or connectors" (descriptors for level 4, mentioned by participant E. She argues that even a level 3 may use appropriate connectors, like "and".
"Speakers at this level will fail to obtain the professional confidence of their interlocutors" (explanation for level 3, mentioned by participant B.	Not clear why the descriptors included "able to speak at length", if pilots and ATCs do not speak at length in RT, mentioned by participant C.
"Levels of fluency will be most apparent during longer utterances in an interaction. They will also be affected by the degree of expectedness of the preceding input which is dependent on familiarity with scripts or schemata" (introductory paragraph for fluency), mentioned by participant B. She argues that "this familiarity with script or schemata has to do with background knowledge, so the degree of expectedness of the preceding input which is dependent on familiarity with scripts or schemata, so is dependent on the knowledge that you have, the technical knowledge that you have on the situation, it is not only the language knowledge that you have. So you will react to a communication, to a prompt depending on the familiarity that it has to your knowledge, to the things that you have inside your brain, so we cannot disregard completely the knowledge, this background knowledge	Unclarity regarding how to rate "able to speak at length", mentioned by participant F. She argued that being "able to speak at length" does not mean you "speak at length". She commented: "if there is a candidate that doesn't speak at length at the start and then occasionally does, you can say they are showing that they are able to even though they didn't do it all the time. Is that what it means?"
"There may be occasional loss of fluency on transition from rehearsed or formulaic speech to spontaneous interaction" (descriptors for level 4), mentioned by participant C. However, he argued that it should have been written more directly, for example, "There may be occasional loss of fluency on switching from standard radiotelephony to plain language".	

Comprehension

Strengths	Weaknesses
"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" (descriptor for level 4), and "failure to understand a clearly communicated unexpected communication even after seeking clarification should result in the assignment of a lower proficiency level assessment" (explanation for level 4), mentioned by participant B. However, she points out that they need clarification as the way they are written now does not account for failure, although the first part of the descriptor says "comprehension is mostly accurate". She argues it is not fair because the descriptors for level 5 say that "comprehension is (…) mostly accurate when the speaker is confronted with a linguistic or situational complication". She suggests that these parts should include adverbs of frequency. Similarly, she points out that the descriptors "may fail to understand a linguistic or situational complication or an unexpected turn of event" (level 3) should also include an adverb of frequency because otherwise anybody who fails to understand something will fail the test, and "we would assign a lot of level 3s".	Difficulty to conceptualize non-standard dialects or regional accents in a language test, mentioned by participant C.
	Difficulty to assess, mentioned by participant A. He argued that it is difficult to assess comprehension "because it is not always immediate that somebody hasn´t quite understood. You listen to the conversation and you think, I wonder if he understood that. And the man will just say "roger". (…)You know then you have to go and test some other way to get his comprehension. So it´s not always immediately obvious, but still very important".

Interactions

Strengths	Weaknesses
"Speakers at this level will not gain the confidence of their interlocutors" (explanation for level 3), mentioned by participant B.	Difficulty to develop test items that will 'introduce mock misunderstandings", mentioned by participant C.
"They display authority in the conduct of the conversation" (level 6 explanation), mentioned by participant E.	Difficulty to rate interactions, mentioned by participants A and C. Participant A argued: "to be honest with you I probably think this is the most difficult to rate, because it talks about the ability to initiate exchanges, to identify and clear up misunderstandings, (…), none of these are totally set apart, a little bit in the link because it is communication, it`s language, here again if something has not been understood has the candidate asked for an explanation? Has he found some other way to understand what the problem is and deal with it?" Participant C believes it is difficult to say the right way to approach how to rate the candidates' ability to deal with misunderstandings.
"Interactions at this level are based on high levels of comprehension and fluency" (explanation for level 5), mentioned by participant D and E.	"Responses are immediate, appropriate and informative (descriptor for level 5), mentioned by participant E. She argues it needs an adverb of frequency because "responses might be appropriate and informative but not immediate all the time".
The adjectives used to explain the responses (immediate, appropriate and informative", mentioned by participant D.	

Appendix I: Participant information sheet

LANCASTER
UNIVERSITY
Department of Linguistics
and English Language

Participant information sheet

Title: *What do ICAO language proficiency test developers and raters have to say about the ICAO language proficiency requirements 12 years after their publication? A qualitative study exploring highly experienced professionals' opinions.*

Researcher: *Angela Carolina de Moraes Garcia*

garcia@lancs.ac.uk / angela.garcia@anac.gov.br

You are invited to take part in this research study. Please take time to read the following information carefully before you decide whether or not you wish to take part.

What is the purpose of this study?

I am carrying out this study as part of my Masters studies in the Department of Linguistics and English Language. The aim of the study is to investigate ICAO language proficiency test developers and raters' perceptions about the ICAO language proficiency requirements in general, its rating scale and the explanation of the descriptors.

What does the study entail?

My study will involve interviews with experienced ICAO test raters and test designers.

Why have I been invited?

I have approached you because you are an experienced ICAO test rater and/or test developer. I would be very grateful if you would agree to take part in my study.

What will happen if I take part?

If you decided to take part, this would involve the following:

I will interview you through Skype in order to learn about your experience with rating and test developing as well as find out your opinion about the ICAO language proficiency requirements. The interview will take approximately 60 minutes. More than one interview session may be necessary. The interview will be audio recorded.

What are the possible benefits from taking part?

Taking part in the interview will allow you to reflect on your own experiences of applying ICAO rating scale or developing test designing a test based on it. If you share your experiences and opinions, your insights will contribute to our understanding of the ICAO rating scale and it might impact on future reviews of the ICAO language proficiency requirements.

What are the possible disadvantages and risks of taking part?

It is unlikely that there will be any major disadvantages to taking part. Taking part will mean investing some time for the interview(s).

What will happen if I decide not to take part or if I don't want to carry on with the study?

Your participation is voluntary. You are free to withdraw from the study at any time and you do not have to give a reason. If you withdraw while the study takes place or until 1 month after it finishes, I will not use any of the information that you provided. If you withdraw later, I will use the information you shared with me for my study.

Will my taking part in this project be kept confidential?

All the information collected about you during the course of the research will be kept strictly confidential. Any identifying information, such as names and personal characteristics, will be anonymised in the Masters dissertation or any other publications of this research. The data I will collect will be securely kept. Any paper-based data will be kept in a locked cupboard. Electronic data will be stored on a password protected computer and files containing personal data will be encrypted.

What will happen to the results of the research study?

The results of the study will be used for academic purposes only. This will include my Masters dissertation and other publications, for example journal articles. I am also planning to present the results of my study at academic conferences. The results of this study will also be informed to ICAO.

What if there is a problem?

If you have any queries or if you are unhappy with anything that happens concerning your participation in the study, please contact myself or my supervisor (Luke Harding – address: County South – Lancaster University – Bailrigg – Lancaster – UK – LA1 4YL – telephone number: +44 1524 593034– e-mail: l.harding@lancaster.ac.uk).

Further information and contact details

garcia@lancs.ac.uk / angela.garcia@anac.gov.br

Thank you very much for considering your participation in this project.

Appendix J: Consent form

LANCASTER
UNIVERSITY
Department of Linguistics
and English Language

Consent Form

Project title: *What do ICAO language proficiency test developers and raters have to say about the ICAO language proficiency requirements 12 years after their publication? A qualitative study exploring highly experienced professionals' opinions.*

1. I have read and had explained to me by *Angela Carolina de Moraes Garcia* the information sheet relating to this project.

2. I have had explained to me the purposes of the project and what will be required of me, and any questions have been answered to my satisfaction. I agree to the arrangements described in the information sheet in so far as they relate to my participation.

3. I understand that my participation is entirely voluntary and that I have the right to withdraw from the project any time, but no longer than 2 months after its completion. If I withdraw after this period, the information I have provided will be used for the project.

4. I understand that all data collected will be anonymised and that my identity will not be revealed at any point.

5. I have received a copy of this consent form and of the accompanying information sheet.

Name:

Signed:

Date: