Defining the construct of aeronautical English listening for pilots: Insights from the literature, the ICAO rating scale, and key stakeholders’ perceptions

by

Angela Carolina de Moraes Garcia

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Applied Linguistics and Discourse Studies

Carleton University

Ottawa, Ontario

© 2023

Angela Carolina de Moraes Garcia
Abstract

In 2003, the International Civil Aviation Organization (ICAO) published its Language Proficiency Requirements (LPRs), requiring pilots and air traffic controllers (ATCOs) to demonstrate their ability to speak and understand the language used in radiotelephony communications. Over the years, multiple interpretations of the policy have led to practices that may have undermined the meaningfulness of the LPRs (ICAO, 2022). This is likely due to issues such as the presence of ambiguity and uncertainty in the policy as well as the lack of clarity regarding the nature of the language that needs to be assessed (e.g., Alderson, 2010; Douglas, 2004; Kim, 2012). Since the ICAO Test Design Guidelines (TDGs) recommend that tests should contain tasks dedicated to the assessment of listening separately from speaking, the goal of this doctoral research was to address the initial stages of the development of a test to assess pilots’ listening comprehension in isolation: the definition of the listening construct. Construct definition is an essential step in test development and validation, and it forms the basis for the development of test specifications. To investigate what should be assessed in such a test, three perspectives were taken into consideration. A review of the academic literature and research was made to better understand the construct of listening tests, and, more specifically, the features of aeronautical English listening that may contribute to misunderstandings between pilots and ATCOs. An analysis of the ICAO rating scale was also conducted to examine the listening construct underpinned by the descriptors. Additionally, the perceptions of pilots and other key stakeholders regarding the assessment needs in relation to the listening performed by pilots in radiotelephony communications with ATCOs was investigated. The conceptual foundation underpinning this research project is the sociocognitive theory developed by Weir (2005), which takes into account both individuals’ cognitive processes and the features of the context in which
the tasks are performed. The results of the three analyses were merged to define the construct based on the operations pilots need to be able to perform and the conditions under which they operate.
Dedication

For my mother, who was my English teacher, and who is always the first to read anything I write.
Acknowledgements

Finalizing a doctoral dissertation is very rewarding. This accomplishment would not have been possible if it was not for the invaluable support and assistance I have received from a number of people and institutions. I feel immensely grateful to everyone who has supported me during this PhD journey. I would not have been able to conduct this research and write this dissertation if I had not received support from my family, my friends, and the community. I would like to show my profound gratitude and appreciation to the following:

- First and foremost, I would like to thank God, as without Him, nothing would have been possible.
- My PhD supervisors, Guillaume Gentil and Angel Arias. The careful and expert guidance they provided was invaluable. I am deeply grateful for the continued support I have received from them. I am profoundly indebted to you. I would also like to thank Geoffrey Pinchbeck, who was my supervisor for a short period of time, but who gave me excellent advice.
- My wonderful advisory committee: Guillaume Gentil, Angel Arias, Janna Fox, Eric Friginal, and Ana Lucia Monteiro for supporting me through this process and for all the time they have spent reading my work, meeting with me, sending me suggestions of articles and books, suggesting the best approaches, and much more. They were incredible, and I will be forever thankful.
- Those who participated in my study by answering the questionnaire or participating in the interviews. I really appreciate their time and willingness to contribute to my research!
The professors who taught the doctoral core seminars in my first year, Janna Fox and Natasha Artemeva. They are very knowledgeable and helped me a lot with the foundation for my PhD research. They also supervised my pilot study, which ended up being a very important part of this dissertation. I owe my greatest gratitude to them.

My cohort classmates, Kathryn Carreau and Gillian Mclellan, with whom I have shared the challenges of the beginning of the program, other graduate students with whom I have learned in classes, workshops, and presentations as well as SALaDS for the amazing work they do.

ANAC, the Brazilian National Civil Aviation Agency, where I work, for giving me the opportunity to have leave of absence and do this PhD. Having conducted this research will help me work towards the improvement of the Brazilian language assessment for civilian pilots.

My colleagues from the Language Proficiency group at ANAC. I would like to especially thank my friend Ana Lucia Monteiro for encouraging me to pursue these academic studies and for being such a great inspiration to me. I would also like to thank my coordinators and managers who authorized me to take a training leave.

Carleton University – I would like to thank all faculty members, especially the graduate supervisor, and the staff, especially Joan Grant, Tracey Wright, and Sarah Henriquez, who were always very supportive. I extend my gratitude to my teaching assistant supervisors Kim David, Don Myles, Helen Triantafellou, Ricardo Tabone, and Alex Myers. I have really enjoyed working for them and learning with them.
- Awards donors and award committees – Thank you very much for helping me financially in the pursuit of my studies.

- The International Civil Aviation English Association (ICAEA), especially the ICAEA Research Group. I would like to give special thanks to Patricia Tosqui-Lucks, Aline Pacheco, Malila Prado, and Ana Lucia Monteiro, with whom I have worked closely in the organization of a webinar on the use of corpus linguistics to teach and assess aeronautical English.

- The Aeronautical English Research Group (GEIA) – I thank the GEIA organizers (especially the coordinator, Patricia Tosqui-Lucks) for putting together annual high-level seminars and for giving us the opportunity to publish the proceedings twice in the past four years. I would also like to thank GEIA members for the opportunity to learn with them through their interesting presentations and constant interactions.

- The International Language Testing Association (ILTA) Aviation English Special Interest Group (SIG) for our fruitful discussions. I would like to especially thank Natalia Andrade and Ana Ligia Silva, with whom I have had frequent valuable meetings. I would also like to thank Maria Treadaway, William Agius, and Ana Lucia Monteiro for their inspiring research, and invaluable feedback.

- The editors and reviewers of the journals and books in which I have published – Thank you for the amazing opportunity to publish my work!

- My family – I would like to thank all my family, but especially my mother, Leda; my father, Roberto; my children, Felipe, Isabela, and David; my partner, Atila; my grandmother, Dalva; my brother, Guilherme; my sisters, Cristina and Fanny; my godmother, Tuca; my niece, Amanda; my cousin, André. They have all supported me
in every way possible from helping me financially in times of need to helping me with my house chores and with taking care of David. They have also given me the emotional support I needed.

- My dog and companion, Sophie, who lay by my side for hours and hours every day while I was writing this dissertation.

- My friends, who encouraged me and were by my side whenever I needed them. I will never forget, for example, how much they have helped me when I had three miscarriages during this PhD.

- My church, who prayed that I would be successful on this PhD journey.

- The doctors, my therapist, and other health professionals who supported me before, during, and after the pandemic.

- My reviewer, Paul Bruthiaux, who helped me immensely with making sure the APA style was properly followed and with the formatting, which is something I consider to be very challenging.

- You, my reader. I have written this dissertation for you. I hope you enjoy reading it and that it helps you in some way.

I apologize for any omissions. I am thankful to anyone who has directly or indirectly helped me in completing this dissertation.
Preface

This manuscript-style dissertation is composed of five manuscripts. The first is a peer-reviewed conference proceedings journal article. The second is a published book chapter. The third and fourth are also published peer-reviewed journal articles (the latter being a co-authored article). They are all reproduced here whole just as they were published. The material follows the original citation style required by each publisher. To access and cite any of these publications, please use the information below.

- Chapter 2


- Chapter 3


- Chapter 4


- Chapter 5

The fifth and last manuscript is a book chapter to be submitted by September 2023. The abstract has already been accepted by the editors and publisher (Bloomsbury). The working title of the book is *Research, Assessment and Pedagogy in Global Aviation English*. The manuscript follows the Chicago Manual of Style, as required by the publisher.

I acknowledge that the materials have been reprinted here with the permission of the publishers. Their permissions were granted as follows:

- Chapter 2 – Permission granted by Prof. Sambor Grucsa on 19 July 2023
- Chapter 3 – Permission granted by Dr. Patricia Tosqui Lucks and Juliana de Castro Santana on 21 July 2023
- Chapter 4 – Permission granted by Prof. Nadežda Stojković on 19 July 2023
- Chapter 5 – Permission granted by André Effgen de Aguiar on 11 July 2023

The publishers of the manuscripts in chapters 2, 3, and 4 adopt the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND) public license (see [https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode](https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode)), and the publisher of the manuscript in Chapter 5 adopts the Creative Commons Attribution 4.0 International (CC-BY 4.0) public license (see [https://creativecommons.org/licenses/by/4.0/](https://creativecommons.org/licenses/by/4.0/)). I acknowledge that there have been no modifications in the published materials.

The co-authored manuscript (Chapter 5) has been reprinted with the permission from the co-author, Dr. Janna Fox, as seen in Appendix A. With assistance from the co-author and as the
lead author, I was fully involved in setting up and conducting the research, obtaining data and analyzing the results as well as preparing and writing up the material presented in the co-authored article integrated in the dissertation (Chapter 5).

A statement from my supervisors can be found in Appendix B in which they confirm that the information I provided in this preface is truthful.
Table of Contents

Chapter 1: Introduction ............................................................................................................. 21
  1. Preparing for takeoff ......................................................................................................... 21
     1.1 Background on the ICAO Language Proficiency Requirements .................................... 23
     1.2 Issues with the ICAO policy ........................................................................................... 28
     1.3 The research need and the overarching research question .............................................. 31
     1.4 Content of this dissertation ............................................................................................. 35
     1.5 Positionality statement .................................................................................................... 39

References .................................................................................................................................... 43

Chapter 2: Manuscript 1 – Investigating the construct of a listening test to assess pilots’
comprehension: A step-by-step project for test developers .................................................... 48

Chapter 3: Manuscript 2 – The listening construct: Theories and implications for the
assessment of pilots and ATCOs................................................................................................ 76

Chapter 4: Manuscript 3 – Investigating the construct of aeronautical English listening
testing: A qualitative analysis of the ICAO rating scale ......................................................... 101

Chapter 5: Manuscript 4 – Contexts and constructs: Implications for the testing of listening
in pilots’ communication with air traffic controllers........................................................................ 121

Chapter 6: Manuscript 5 – The construct of aviation English listening tests for pilots from
three perspectives: Literature, policy, and perceptions of pilots and other key stakeholders
....................................................................................................................................................... 154
1. Introduction ........................................................................................................................... 156

2. The aeronautical English listening construct according to the literature ....................... 158
   2.1 Frameworks for the assessment of listening comprehension ........................................ 159
   2.2 Listening to aeronautical radiotelephony communications ........................................... 163

3. The listening construct in the ICAO rating scale ................................................................. 167

4. The aeronautical English listening construct and perceptions of key stakeholders ......... 172

5. A description of the aeronautical English listening construct from three perspectives ...... 177

6. Conclusion ............................................................................................................................ 180

References .................................................................................................................................... 182

Chapter 7: Conclusion ................................................................................................................ 187

7.1 Final approach ..................................................................................................................... 187

7.2 Limitations .......................................................................................................................... 188

7.3 Next flights? ....................................................................................................................... 189

7.4 Suggestions of further research ....................................................................................... 191

7.5 Landing ............................................................................................................................... 193

References .................................................................................................................................... 195

Appendix A: Co-author Permission Letter .............................................................................. 199

Appendix B: Statement from supervisors ............................................................................... 200

Appendix C: Supervisor Review and Approval Form (Chapter 5) ......................................... 201
Appendix D: Tri-Council Policy Statement 2 (Course Certificate) ........................................ 202

Appendix E: Email invitation (Chapter 5, Phase 1) ............................................................... 203

Appendix F: Online advertisement (Chapter 5, Phase 1) ...................................................... 204

Appendix G: Email Invitation (Chapter 5, Phase 2) .............................................................. 205

Appendix H: Online Advertisement (Chapter 5, Phase 2) ...................................................... 207

Appendix I: Informed Consent Form (Chapter 5, Phase 1) .................................................... 209

Appendix J: Informed Consent Form (Chapter 5, Phase 2) .................................................... 212

Appendix K: Interview Guide (Chapter 5, Phase 2) .............................................................. 217

Appendix L: Ethics Clearance Certificate (Chapter 5) ............................................................ 220
# List of Tables

## Chapter 1

**Table 1.1** The ICAO Holistic Descriptors (ICAO, 2022, p. APP 1-1) .............................24

**Table 1.2** The ICAO Language Proficiency Rating Scale (ICAO, 2022, p. ATT A-1) ........25

## Chapter 2

**Table 1.** Specific test development questions (STDQ) ....................................................53

**Table 2.** Sphere of construct, as defined by U. Knoch/ S. Macqueen (2020) related to STDQ 2 to 5 .......................................................................................................................56

**Table 3.** The areas of needs analysis for STDQ 1 to 8 .......................................................58

**Table 4.** Information that might be included in the Design Statement ..............................59

**Table 5.** Type of data and sources of data per STDQ ..........................................................62

**Table 6.** Participants in the study per STDQ .................................................................63

**Table 7.** Examples of questions that may be asked in the interview with the test developer/ requester ..................................................................................................................63

**Table 8.** Literature review to be conducted in Step 2 (STDQ 3) ........................................65

**Table 9.** Examples of questions that may be asked in the interview with the recognized test developer ......................................................................................................................66

**Table 10.** Examples of questions that may be asked in the questionnaire .........................68

**Table 11.** Examples of questions that may be asked in the focus group .............................69

## Chapter 4

**Table 1.** Comprehension differences among levels 3 to 5 ..................................................110
Table 2. Listening construct represented in the comprehension descriptors.........................112

Table 3. Elements test developers may consider when developing a listening test to assess pilots and ATCOs’ listening in isolation.................................................................115

Chapter 5

Table 1. Participants’ experience with international flights.................................131

Table 2. Reported final levels awarded to participants on aeronautical English language proficiency tests..............................................................131

Table 3. How important is it to assess the ability to understand speech at different speeds?........133

Table 4. How frequently do pilots hear ATCOs speaking at different speech rates?...............133

Table 5. How frequently do pilots hear unintelligible pronunciation in RT communications?....135

Table 6. How important is it to assess the ability to understand different accents?...............135

Table 7. How frequently, in RT communications, do pilots hear words and expressions whose meaning they do not understand?...............................................................136

Table 8. How important is it to assess the ability to understand the meaning of uncommon words and expressions?..................................................................................................136

Table 9. How important is it to assess the ability to recognize indirect meaning?...............138

Table 10. What was the maximum number of pieces of information participants listened to in a single transmission from an ATCO?.............................................................................139

Table 11. How important is it to assess comprehension of plain English and phraseology?.....140

Table 12. How frequently, in RT communications, do pilots hear sentences they cannot understand because of the way the words are put together (structure)?..............................................141
Table 13. How important is it to assess the ability to understand main ideas, the ability to recognize words and numbers and the ability to extract specific meaning from both short and long transmissions?..............................................................................................................................142

Table 14. Summary of pilots’ listening comprehension needs and implications for testing……146

Chapter 6

Table 1. What listening involves according to the literature..............................................162

Table 2. Definition of the construct of aeronautical English listening test for pilots.............180
List of Figures

Chapter 2

Figure 1. Overview of proposed project.................................................................54

Figure 2. LAPP test development cycle, adapted from U. Knoch/ S. Macqueen 2020: 97……57

Figure 3. Step 1 Diagram (STDQ 1).................................................................64

Figure 4. Step 2 Diagram (STDQ 2 and 3)..........................................................64

Figure 5. Step 3 Diagram (STDQ 4).................................................................65

Figure 6. The Explanatory Sequential Design followed in the pilot study...............67

Figure 7. An Explanatory Sequential Design to be followed in Step 5 (adapted from J. Creswell/ V.Clark 2018: 85).................................................................70

Chapter 6

Figure 1. A summary of what is involved in listening according to the literature.........163

Figure 2. Features of aeronautical English listening that may contribute to miscommunications.................................................................164

Figure 3. The four spheres of the listening construct represented in the ICAO rating scale.....171

Figure 4. Summary of the aeronautical English listening construct from three perspectives….178

Figure 5. Abilities that should be assessed to a lesser extent........................................179
List of Acronyms

ANAC – Agência Nacional de Aviação Civil (National Civil Aviation Agency – Brazil)
AI – Artificial Intelligence
ATC – Air Traffic Control
ATCO – Air Traffic Control Officer (Air Traffic Controller)
ATIS – Automatic Terminal Information Service
CUREB – Carleton University Research Ethics Board
DOC 9835 – Manual on the Implementation of ICAO Language Proficiency Requirements
EL2 – English as a Second Language
ELE – English Language Expert
EUROCONTROL – European Organisation for the Safety of Air Navigation
FAA – Federal Aviation Administration
GEIA – Grupo de Estudos em Inglês Aeronáutico (Aeronautical English Research Group)
ICAEA – International Civil Aviation English Association
ICEA – Instituto de Controle do Espaço Aéreo (Airspace Control Institute – Brazil)
ICAO – International Civil Aviation Organization
ILTA – International Language Testing Association
LAASIG – Language Assessment in Aviation Special Interest Group
LAPP – Language Assessment for Professional Purposes
LPR – Language Proficiency Requirement
LSP – Language for Specific Purposes
LTRC – Language Testing Research Colloquium
NES – Native English speaker
NNES – Non-native English speakers

PRICESG – Proficiency Requirements in Common English Study Group

QUAL/qual – Qualitative research (uppercase: primary research approach; lowercase: secondary research approach)

QUAN/quan – Quantitative research (uppercase: primary research approach; lowercase: secondary research approach)

RQ – Research Question

RT – Radiotelephony

SARP – Standards and Recommended Practices

SDEA – Santos Dumont English Assessment

SME – Subject Matter Expert

STDQ – Specific Test Development Question

TCPS – Tri-Council Policy Statement

TDG – Test Design Guidelines

TLU – Target Language Use
Chapter 1: Introduction

1. Preparing for takeoff

The oldest language test I know of is described in the Bible, in the Book of Judges, Chapter 12. About 1,100 years before Christ, Jephthah, from the region of Gilead, was a judge in Israel for six years. He fought a battle against the Ephraimites and won. When crossing the Jordan river to escape, those who survived the battle were asked by the men of Gilead if they were Ephraimites. Obviously, they would answer that they were not. Then, they would be requested to say the word *Shibboleth*. According to the Bible, people from Ephraim would pronounce this word with an *s* sound (*Sibboleth* instead of *Shibboleth*). Thus, the Gileadites were able to identify those who were from Ephraim and killed them in the Jordan river. The story claims that 42,000 Ephraimites died at that time. Passing this “language test” was a matter of life or death.

Shibboleth-like tests have been recently used in conflict situations, such as the civil war in Lebanon in 1976 and the civil war in Sri Lanka in 1983 as well as to detect unauthorized immigrants (McNamara, 2005). Nowadays, the consequences of failing a language proficiency test might not be as deadly, but they may still have a significant impact on the test takers and on society in general. A case in point is aeronautical English testing. Let us imagine that a pilot or an ATCO (air traffic controller) passes a language proficiency test without being proficient in aeronautical English (we will go into more detail about what aeronautical English is later in this chapter), and this lack of proficiency contributes to an accident that kills hundreds of people. On the other hand, if the pilot or ATCO does not pass the test, other consequences may arise, including potentially affecting the test taker’s career and welfare as well as their family’s. McNamara (2005) argues that even ordinary language proficiency tests, which are created to
establish proficiency in communicative or educational contexts, may assume a Shibboleth-like function as they may serve as “an instrument of public policy in contexts of intergroup competition” (p. 364). He adds that “all language testing is potentially political; it can be associated with, and operate in the service of, power and control” (p. 368). This dissertation addresses a language testing situation embedded in a highly political context: the assessment of pilots and ATCO language proficiency in accordance with the International Civil Aviation Association (ICAO) Language Proficiency requirements (LPR).

This doctoral dissertation is a manuscript-style dissertation, what Carleton University refers to as “an integrated thesis” (Carleton University, n.d.). As Anderson et al. (2021) define it:

A manuscript-style macrostructure is comprised of either published or publishable manuscripts, usually in the style of journal articles or book chapters, and often includes separate introductory and concluding chapters. This format can vary in length, involving only one or multiple interior manuscript-chapters as well as authorship status, with manuscripts being either single or co-authored texts. (p. 3)

Manuscript-style dissertations are becoming increasingly popular as they offer various advantages, including helping to promote academic production and to propagate doctoral research more easily and widely (Anderson et al., 2021). This dissertation style aligned with my goals because I was able to disseminate my findings much faster than if I were to write a traditional dissertation. In addition, I was able to dedicate time and energy to publishing since what I was writing would be used for my dissertation. The downside of this type of dissertation is that there are usually overlaps in the chapters, usually in the background and literature review sections.
This introductory chapter starts with background information on the ICAO LPRs. Then, I introduce the main discussions that have featured in the literature since the publication of the ICAO LPRs (including issues related to power and fairness, as mentioned), indicate the research need, state the purpose of this dissertation and the overarching research question that guided this research, and define some important terms. Next, I indicate the content of this dissertation and provide essential information about each manuscript included in it. Finally, I present my positionality as a researcher by describing my personal values, beliefs, history, and experience.

1.1 Background on the ICAO Language Proficiency Requirements

In 2003, in response to India’s request to strengthen the provisions for pilots and ATCOs’ language proficiency (Popa, 2019), the ICAO Council adopted Amendment 164 to Annex 1 (Personnel Licensing) of the Convention on International Civil Aviation (ICAO, 2022a). India made this request after a catastrophic accident occurred in its airspace in 1996 when two aircraft collided mid-air, resulting in 349 fatalities. The main cause of the accident was said to be the lack of English proficiency of one of the flight crews. The ICAO LPRs state that “aeroplane, airship, helicopter, and powered-lift pilots; aeroplane, airship, glider, rotorcraft, powered-lift, or free balloon remote pilots; air traffic controllers; and aeronautical station operators shall demonstrate the ability to speak and understand the language used for radiotelephony communications to the level specified in the language proficiency requirements in Appendix 1” (pp. 1-16). According to Appendix 1:

To meet the language proficiency requirements … an applicant for a license or a license holder shall demonstrate, in a manner acceptable to the Licensing Authority, compliance with the holistic descriptors at Section 2 and with the ICAO Operational Level (Level 4)
of the ICAO Language Proficiency Rating Scale in Attachment A. (ICAO, 2022a, p. APP 1-1)

The set of five holistic descriptors (shown in Table 1.1) stipulates what proficient pilots and ATCOs are expected to be able to do as a whole and establishes the context in which communication occurs.

**Table 1.1**

*The ICAO Holistic Descriptors* (ICAO, 2022a, p. APP 1-1)

<table>
<thead>
<tr>
<th>Proficient speakers shall:</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) communicate effectively in voice-only (telephone/radiotelephone) and in face-to-face situations;</td>
</tr>
<tr>
<td>b) communicate on common, concrete and work-related topics with accuracy and clarity;</td>
</tr>
<tr>
<td>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;</td>
</tr>
<tr>
<td>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</td>
</tr>
<tr>
<td>e) use a dialect or accent which is intelligible to the aeronautical community.</td>
</tr>
</tbody>
</table>

The ICAO Language Proficiency Rating Scale seen in Table 1.2 is divided into six analytic criteria (pronunciation, structure, vocabulary, fluency, comprehension, and interactions) and describes the discrete features of language use at six proficiency levels (level 1, Pre-
elementary; level 2, Elementary; level 3, Pre-operational; level 4, Operational; level 5, Extended; and level 6, Expert). The ICAO rating scale is a pivotal part of the ICAO LPRs as it is used to rate pilots’ and ATCOs’ proficiency worldwide independently of the test they take, and it was the basis for the development of tests by Contracting States and testing service providers.

Table 1.2

The ICAO Language Proficiency Rating Scale (ICAO, 2022a, p. ATT A-1)

<table>
<thead>
<tr>
<th>LEVEL</th>
<th>PRONUNCIATION</th>
<th>STRUCTURE</th>
<th>FLUENCY</th>
<th>COMPREHENSION</th>
<th>INTERACTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert 6</td>
<td>Pronunciation, stress, rhythm, and intonation, though possibly influenced by the first language or regional variation, almost never interfere with ease of understanding.</td>
<td>Both basic and complex grammatical structures and sentence patterns are consistently well controlled.</td>
<td>Vocabulary range and accuracy are sufficient to communicate effectively on a variety of familiar and unfamiliar topics. Vocabulary is idiomatic, nuanced, and sensitive to register.</td>
<td>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.</td>
<td>Comprehension is consistent 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.</td>
</tr>
<tr>
<td>Extended 5</td>
<td>Pronunciation, stress, rhythm, and intonation, though influenced by the first language or regional variation, rarely interfere with ease of understanding.</td>
<td>Basic grammatical structures and sentence patterns are consistently well controlled. Complex structures are attempted but with errors which sometimes interfere with meaning.</td>
<td>Vocabulary range and accuracy are sufficient to communicate effectively on common, concrete, and work-related topics. Paraphrases consistently and successfully. Vocabulary is sometimes idiomatic.</td>
<td>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.</td>
<td>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.</td>
</tr>
<tr>
<td>Operational 4</td>
<td>Pronunciation, stress, rhythm, and intonation are influenced by the first language or regional variation but sometimes interfere</td>
<td>Basic grammatical structures and sentence patterns are used creatively and are usually well</td>
<td>Vocabulary range and accuracy are usually sufficient to communicate effectively on rehearsed or occasion.</td>
<td>Produces stretches of language at an appropriate tempo. There may be occasional loss of fluency on transition from rehearsed or used is sufficiently</td>
<td>Comprehension is mostly accurate on common, concrete, and work-related topics when the accent or variety used is sufficiently</td>
</tr>
</tbody>
</table>
with ease of understanding. Errors may occur, particularly in unusual or unexpected circumstances, but rarely interfere with meaning.

### Pre-operational 3

- Pronunciation, stress, rhythm, and intonation are heavily influenced by the first language or regional variation and usually 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. Interaction is limited to simple routine exchanges. 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.
- Performs at a level below the Elementary level.

### 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. Interaction is limited to simple routine exchanges. Response time is slow and often inappropriate.
- Performs at a level below the Elementary level.

### Pre-elementary 1

- Performs at a level below the Elementary level.

---

According to the ICAO LPRs, for a pilot to fly internationally and for a controller to control international flights, they need to perform at a minimum of level 4 in all six criteria, as the test taker’s final level corresponds to the lowest level achieved. Pilots and ATCOs who demonstrate language proficiency at level 4 should be evaluated at least once every three years, and those who demonstrate language proficiency at level 5 should be evaluated at least once every six years. Those who demonstrate language proficiency at level 6 do not need to be
evaluated again. Moreover, native or native-like speakers do not ever need to be formally evaluated.

A note to Annex 1 adds some relevant information. It explains that the language the ICAO LPRs refer to is either English or the language normally used by the station on the ground. In practice, the language that is commonly evaluated is English because, when pilots and ATCOs use the language used by the station on the ground, they usually use their native language, and, as seen, they do not need to be formally evaluated in this case.

To support Contracting States in the implementation of the ICAO LPRs, in addition to Annex 1, ICAO has published the following materials:


ICAO is going to publish the ICAO Test Design Guidelines (ICAEAa, n.d.) soon in the form of a handout.

Although the ICAO LPRs were published in 2003, they would only become applicable on March 5, 2008. To help Contracting States that had been unable to comply with the requirements by 2008, ICAO extended this deadline to March 5, 2011. In 2011, ICAO allowed Contracting States that could not comply with the language proficiency requirements by that date “to send to ICAO their implementation plans, which were to be updated on a regular basis after March 2011 until full compliance was reached” (ICAO, 2013, p. 2). ICAO reported that as of January 3,
2013, the rate of effective implementation of the ICAO LPRs among Contracting States was 67.75%.

1.2 Issues with the ICAO policy

Since the publication of the ICAO LPRs, issues have been identified in the policy in general and, more specifically, in the holistic descriptors and the ICAO language proficiency rating scale (e.g., Alderson 2009, 2010; Douglas, 2004, 2014; Kim, 2012; Knoch, 2009, 2014). One of the main problems discussed in the literature is the burden the policy puts on non-native speakers of English (NNES). Although ICAO states in the second edition of DOC 9835 that “the burden for improved communications should not be seen as falling solely on non-native speakers,” the ICAO policy does not require native (NES) or native-like speakers to be formally assessed. The policy seems to ignore that miscommunications also occur between native speakers of English (e.g., Prinzo et al., 2008; Tajima, 2004). Additionally, communicating effectively in radiotelephony communications encompasses more than just being proficient in a language. Native speakers need to make an effort to understand non-native speakers and to be understood. Therefore, many authors have argued that native speakers should also be formally assessed (Borowska, 2017; Douglas, 2014; Estival et al., 2016; Garcia, 2015; Kim, 2012; McNamara et al., 2019; Monteiro, 2019; Read & Knoch, 2009; Trippe, 2018). As discussed earlier in this chapter, language testing is a political activity, which involves social values and leads to social consequences (McNamara, 2005). By having only non-native speakers be formally evaluated, the ICAO policy may not only be ignoring a crucial feature of this assessment context but also contributing to the power, interests, and control of a limited number of nations.
Another issue that is recurrently discussed in the literature related to the ICAO LPRs is the lack of clarity, sometimes even the presence of ambiguity, in the construct of the ICAO LPRs (e.g., Douglas, 2004). The construct can be understood in simple terms as “the thing we are trying to measure” (Buck, 2001, p. 1). This definition is not only simple, but it also implies (given the use of the verb *try*) that a test may be measuring something the test does not intend to measure. As Knoch and Macqueen (2020) argue, a test may measure something that is different from what it states to measure or even from what stakeholders perceive the test is measuring. Knoch and Macqueen propose there are four dimensions of the construct: (i) the stated construct (the construct, for example, stated by the testing service provider on the test website, or by the policy/rating scale); (ii) the operationalized construct (the construct actually tested through the test, as seen, for example, through test takers’ performance in the test or in the test specifications and test versions); (iii) the theoretical construct, which they define as “the theoretical, underlying, unobservable ability which is assumed to explain differences in assessee responses and their classification (e.g., score) differences” (p. 49); and (iv) the perceived construct (what the test takers and other stakeholders believe the test is assessing). Examples of this lack of clarity and ambiguity of the ICAO policy include:

- A note to Appendix 1 explains that “the language proficiency requirements are applicable to the use of both phraseologies and plain language” (ICAO 2022a, p. APP 1-1). Based on this wording, test developers might understand that both phraseology and plain language should be assessed. However, DOC 9835 states that “it is acceptable that a test contains a scripted task in which phraseology is included in a prompt, but the test should not be designed to assess phraseology” (ICAO, 2010, pp. 6-12).
• DOC 9835 (ICAO, 2010) allows test developers to interpret the construct in two different ways. It states that:

It is important to keep in mind that the idea of a work-related context can accommodate different interpretations. A narrow interpretation would aim to closely replicate radiotelephony communications, including the extent of plain language needed in unusual, unexpected or emergency situations. A broad interpretation of the holistic descriptors and Rating Scale would aim to elicit plain language on various topics that are related to radiotelephony communications or aviation operations, without replicating radiotelephony communications specifically. Examples may include question and answer routines, problem-solving exchanges, briefings, simulations and role-plays. Both interpretations are valid. (pp. 6-7)

The possibility of having different interpretations of the construct may have contributed to practices that might have undermined the meaningfulness of aviation English tests, as recently acknowledged by ICAO (ICAO, 2022b).

• According to the holistic descriptors, “proficient speakers shall communicate effectively in voice-only (telephone/radiotelephone) and in face-to-face situations” (ICAO, 2022a, p. APP 1-1). However, DOC 9835 states that the purpose of the ICAO LPRs is to assess pilots and ATCOs’ proficiency in the language of aeronautical radiotelephony communications. This kind of communication is voice-only: there is no face-to-face communication. The fact that the holistic descriptors include face-to-face communications might refer to communications between pilots in the cockpit or between an ATCO and their supervisor, but this is not clear.
Soon after the publication of the ICAO LPRs, Douglas (2004) argued that:

It will be necessary to gather extensive and detailed information about the nature of 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. Such information will form input for the definition of the construct of aviation English that will be measured by the new tests, and which will have to be articulated with the existing Holistic Descriptors and the Language Proficiency Rating Scale that are described in the revised requirements. Areas of ambiguity and uncertainty will need to be clarified, and assessment criteria specified that reflect the purpose of the test, the construct definition, and features of the descriptors and rating scale. (p. 251)

Alderson (2009) also argued more research was needed (e.g., on the nature of the language used in non-normal aviation communications, the ICAO rating scale, and the politics involved in the development and implementation of the policy). This dissertation aims to fill in one of these research gaps, which is to better understand the listening construct of pilots’ aeronautical English tests.

1.3 The research need and the overarching research question

As cited in Section 1.1, ICAO is soon going to publish a handbook with the test design guidelines (TDGs), which were developed in 2019 by the International Civil Aviation English Association (ICAEA), in partnership with ICAO. The TDGs comprise eight criteria that aim to assist Civil Aviation Authorities (CAAs) and test developers to understand key issues related to language test design. Criterion number 3 states that “test instruments need to contain tasks dedicated to assessing listening comprehension, separate from tasks designed to assess speaking performance” (ICAEAA, 2023, Criterion 3). They explain that “tests need to contain sections and
parts which are designed to only assess listening comprehension” (ICAEAa, 2023, Criterion 3).

The reason why the TDGs justify the need to assess listening in isolation is the following:

Listening comprehension represents at least half the communicative load in aeronautical communication. Proficiency in comprehension is determined by a range of different cognitive skills, language skills and knowledge. All of these attributes are internal and cannot be directly observed for assessment purposes. In contrast, speaking skills are more observable and can be assessed directly by observing speaking performance. Therefore, proficiency in listening comprehension is best assessed in contexts which are not affected by speaking ability because basing decisions on what test-takers say may be more of a result of their speaking skills rather than their internal comprehension proficiency.

(ICAEAb, 2023, What does this mean for test design?)

As seen in this explanation, the TDGs consider that when assessing listening only together with speaking, skills that are irrelevant to the construct may play a major role in the assessment process and influence test results. This inadequate operationalization of the construct by including in the test the assessment of abilities that are not important to the construct is known as construct irrelevant variance (Messick, 1989). and is a threat to validity. Another threat to validity pointed out by Messick is when a test omits important dimensions of the construct, and is called construct underrepresentation.

There are other reasons why pilots’ listening should also be assessed in isolation, which are not mentioned by the ICAO test design guidelines. Although most of the listening performed by pilots in their communication with ATCOs takes place as part of an interaction, which is why interactive listening should be assessed, pilots also need to listen to recordings in isolation, including Automatic Terminal Information Service (ATIS) messages. ATIS messages, which are
usually recorded, provide current information about the airport to departing and arriving aircraft. These recordings include information about the weather (wind direction and speed, visibility, temperature, dew point, altimeter setting, etc.), active runways, airport conditions, radio frequencies, and other information about the airport and its vicinity.

Another reason why assessing pilots’ listening in isolation is not only important but highly compelling is that pilots need to maintain situational awareness by listening to communications between other pilots and the air traffic control. Situational awareness can be defined as:

The accurate perception and understanding of all the factors and conditions within the five fundamental risk elements (flight, pilot, aircraft, environment, and type of operation) that comprise any given aviation situation and affect safety before, during, and after the flight. (FAA, 2016, p. 2-24)

It is therefore crucial for pilots to monitor radiotelephony communications for traffic, weather information, and air traffic control interactions in order “to enhance situational awareness by helping the pilot develop a mental picture of what is happening” (FAA, 2016, p. 2-24).

To summarize, it is important to assess pilots’ listening in isolation for four main reasons:

1) Criterion 3 of the ICAO test design guidelines recommends that listening should be assessed separately from speaking;

2) To minimize construct irrelevant variance;

3) Because pilots need to listen to recorded aeronautical information such as ATIS; and

4) Because pilots need to monitor radio communications in order to maintain situational awareness.
Thus, the purpose of this dissertation is to address the initial stage of a test (or part of a
test) that assesses pilots’ listening in isolation: the definition of the construct. Since the
publication of the ICAO LPRs, there have been some studies on the construct of aeronautical
English assessment. Kim (2012), for example, explored the construct of radiotelephony
communications from the perspective of Korean pilots and ATCOs. Garcia (2015) investigated
the perceptions of experienced test developers and raters regarding the ICAO LPRs and the
language proficiency rating scale. Monteiro (2019) examined and specified the proficiency
construct in intercultural radiotelephony communications between pilots and ATCOs. Silva
(2022) investigated the construct of a test that aims to assess the language proficiency of military
pilots who fly in the Air Demonstration Squadron of the Brazilian Air Force. Together, these
studies have contributed to improve our understanding on the general aeronautical English
construct and the need to revisit the ICAO LPRs. However, none of these studies specifically
addressed the construct of a listening test in isolation. In response, this dissertation addresses the
following overarching research question:

*How can the aeronautical English construct be defined in a test that aims to assess
pilots’ listening in isolation?*

To address this overarching research question, three perspectives were adopted. The
purpose is to define a single construct, namely the construct of pilots’ aeronautical listening test,
based on multiple perspectives. First, I investigated what the literature says about the assessment
of listening and the nature of aeronautical English (Garcia, 2022). I then analyzed the ICAO
rating scale to better understand the listening construct underpinned in the policy (Garcia, 2023).
Finally, I jointly conducted an empirical investigation to explore pilots’ and other key
stakeholders’ perceptions on the nature of the listening tasks performed by pilots (Garcia & Fox, 2020).

In this research question and in this dissertation as a whole, the term aeronautical English is preferred over the traditionally used term aviation English, as recommended by Tosqui-Lucks and Silva (2020). Aviation English is a broader term that may refer to any Englishes that are used in the aviation context, which include the English used by pilots and ATCOs in their radiotelephony communication (which they refer to as “aeronautical English”) but also other aviation-related Englishes used by other professionals, such as engineers, technicians, meteorologists, mechanics, flight attendants, even pilots and ATCOs outside the radiotelephony context (cockpit communications, pilots to flight attendants, pilots to passengers, ATCOs to supervisors, etc.). As Tosqui-Lucks and Santana (2022) point out, “other researchers from the international community have adopted this distinction in order to avoid misunderstanding” (p. 10).

1.4 Content of this dissertation

This manuscript-style dissertation is composed of this introduction, five manuscripts, a conclusion, and appendices. I have chosen not to include a Methods chapter because the methods have been explained in the data chapters. To guarantee replicability, the detailed descriptions of the procedures that were conducted which are not in the data chapters are included in the appendices and referred to in this subsection. Inevitably, the content of the chapters may overlap, especially when providing background information.

Chapter 2 consists of a conference proceedings article that has been published in the journal Applied Linguistics Papers (University of Warsaw). The title of the article is “Investigating the construct of a listening test to assess pilots’ comprehension: a step-by-step
project for test developers.” These conference proceedings result from a presentation given during the 9th Aeronautical English Research Group (GEIA) Seminar held in October 2022. The purpose of the presentation was to introduce my proposal for this dissertation, which I had defended earlier that year. Although the project was approved by my advisory committee, it would take a very long time for me to conduct the proposed research and graduate. It was therefore suggested that I could write a manuscript-style dissertation instead of a traditional dissertation. Thus, I decided that instead of submitting a paper explaining my former doctoral research proposal, I would adapt the proposal and submit a paper illustrating a construct definition project that could be followed by test developers. This article is included in this dissertation because it provides a theoretical rationale for a construct definition project. The proposed project may inspire not only aeronautical English test developers but all those who are interested in developing tests to assess language for professional purposes.

The manuscript in Chapter 3 consists of a chapter in an e-book edited by Tosqui-Lucks and Santana published in 2022 as the proceedings of the 8th GEIA Seminar, which took place in November 2021. This chapter, entitled “The listening construct: Theories and implications for the assessment of pilots and ATCOs,” provides a comprehensive review of the literature on what listening involves and how it is assessed with a view to drawing implications for the assessment of pilots’ listening in the context of radiotelephony communications. The chapter addresses the following question:

*What main theories have informed the definition of the listening construct in language testing and assessment, and what are some of their implications for the testing of aeronautical English listening?*
The article reprinted in Chapter 4, entitled “Investigating the construct of aeronautical English listening testing: A qualitative analysis of the ICAO rating scale,” was published in the *Journal of Teaching English for Specific and Academic Purposes* in April of 2023. The purpose of the article was to investigate the listening construct in the comprehension descriptors of the ICAO rating scale as well as how the descriptors for the other criteria in the rating scale could inform the definition of the listening construct. The study addressed the following research questions:

1: *What is the listening in isolation construct represented in the descriptors for comprehension detailed in the ICAO Language Proficiency Rating Scale?*

2: *How may the descriptors for the other criteria help to define the construct of a test to assess listening in isolation?*

Chapter 5 consists of a journal article published in *The ESPracticalist* journal. The title of the article is “Contexts and constructs: Implications for the testing of listening in pilots’ communication with air traffic controllers.” The goal of the study was to investigate pilots’ and other key stakeholders’ perceptions on the nature of the listening tasks performed by pilots as an essential step in construct definition, test development, and validation. This mixed-methods study applied a two-phase explanatory sequential research design as defined by Creswell (2015). It addressed the following research questions:

1: *Based on accounts of pilots and other key stakeholders, what needs to be assessed in relation to the listening performed by pilots in radiotelephony communications with the ATCOs?*

2: *Based on accounts of pilots and other key stakeholders, what are some of the implications of the perceived assessment needs for the development of listening tasks?*
In Phase 1, 156 pilots answered a questionnaire on their perceptions about what a listening test should assess, while in Phase 2, six aeronautical English experts were interviewed. The experts were four experienced pilots who work as raters (one of them also works as an aeronautical English researcher), one ATCO who also works as a rater, and one experienced test developer, rater trainer, and aeronautical English researcher. Since this study involved contact with human participants, it required clearance from Carleton University’s Research Ethics Board – A (CUREB-A). Extensive documentation was therefore submitted, including:

- Supervisor review and approval form (see Appendix C)
- Tri-Council Policy Statement 2 (TCPS 2) course certificate (see Appendix D)
- Recruitment materials:
  - Email invitation for Phase 1 (see Appendix E)
  - Online advertisement for Phase 1 (see Appendix F)
  - Email invitation for Phase 2 (see Appendix G); and
  - Online advertisement for Phase 2 (see Appendix H)
- Informed consent forms:
  - Informed consent form for Phase 1 (see Appendix I); and
  - Informed consent form for Phase 2 (see Appendix J)
- Data collection instruments
  - Questionnaire for Phase 1, shown in the Attachment to the published article (Chapter 5); and
  - Interview guide (see updated interview guide in Appendix K).

The board granted ethics clearance on November 19, 2019 (see Ethics Clearance Certificate in Appendix L).
Chapter 6 consists of a book chapter entitled “The construct of an aeronautical English listening test for pilots from three perspectives: Literature, policy, and perceptions of pilots and other key stakeholders.” The abstract has been accepted by the editors and the publisher (Bloomsbury). This chapter acts as the discussion part of the dissertation and brings together the three perspectives discussed in Chapters 3, 4, and 5. It concludes with a definition of the pilots’ listening test construct.

This dissertation ends with a conclusion chapter that reminds the reader of the purpose of this dissertation and its contribution to the field. This final chapter also identifies the limitations of the present research, explains the next steps that can be followed in the development of a listening test for pilots, and suggests potential areas of further research. Brief concluding remarks are given to close the dissertation.

1.5 Positionality statement

Jacobson and Mustafa (2019) argue that:

The way that we as researchers view and interpret our social worlds is impacted by where, when, and how we are socially located and in what society. The position from which we see the world around us impacts our research interests, how we approach the research and participants, the questions we ask, and how we interpret the data. (p. 1)

Given that this dissertation is mostly qualitative in nature, it is critical to acknowledge and disclose my identity and positionality within this research by reflecting on how my worldview and experience may have impacted the research process.

First of all, I am a non-native speaker of English. I am from Brazil, and my first language is Portuguese. I started learning English as a foreign language when I was nine years old. My mother, who is an English teacher, opened an English school in our house in 1991 and taught
family members and neighbors. Her business grew fast, so I used to listen to English being taught in our house all day. I soon learned the importance of teaching and learning English. When I graduated from high school, I started teaching English to children and teenagers in my mother’s school. I taught English during my undergraduate studies (I graduated in Education). While teaching, I learned a great deal about language, teaching, and testing English in the classroom. I also learned about students’ difficulties and motivations. This personal experience has helped me understand the position of non-native speakers of English in the communication process as well as the amount of effort non-native speakers need to make to become effective communicators in a foreign language.

When I was 25 years old, I started working for the Brazilian National Civil Aviation Agency (ANAC) as a Civil Aviation Regulation Specialist. Since then, I have been working on the implementation of the ICAO LPRs for pilots in Brazil. Our team has developed a test named *Santos Dumont English Assessment* (SDEA), which aims to assess Brazilian civilian pilots’ proficiency in aeronautical English. In my first eight days at ANAC, I took the SDEA examiner course. At that point, I knew nothing about pilot-ATCO communication or the ICAO LPRs. By the end of the eight-day course, I was already creating test items and test versions, and working as a rater. Soon, I would be working with rater training, test item trials, test monitoring, and validation research, among other activities. The fact that I work as a professional in the field of aeronautical English assessment has had a great influence on my purpose for conducting research. I conduct research not only to learn more about the topics under investigation but also to be able to put the research findings into practice. That is why my research questions address an issue that needs to be investigated for the benefit of the ANAC testing process (a social
I would like to add that I am a Christian, and being a Christian also impacts the way I conduct research. Colossians Chapter 3 verse 23 in the Bible says, “whatever you do, work at it with all your heart, as working for the Lord, not for human masters.” I felt I could not continue to work in such a high-stakes testing context without a sound theoretical foundation. I needed to learn more about language assessment in order to work the best way I could. Fairness is also a Christian value I possess. I wanted to ensure that both the test design and the rating process, in which I would act as a rater and rater trainer, were fair. Thus, in 2013, I decided to do a masters in Language Testing at Lancaster University. Studying for this degree broadened my mind and made me understand that the construct underpinned by the ICAO policy needed to change, and that the SDEA needed to be improved. I wrote a working paper that was accepted by ANAC and presented by Brazil to the ICAO council, inviting the ICAO Assembly to revise the ICAO LPRs (ANAC, 2016). I also wrote the first draft of important documents, including the SDEA Task Test Specifications and Blueprint, and the SDEA Guidelines for Interlocutor Behavior, and designed an approach to monitoring both inter and intra-rater reliability. Additionally, I conducted research to investigate the perceptions of test takers in relation to the SDEA and its impact on safety and teaching (Garcia, 2017). This shows that before starting this doctoral research, I was already very much involved in the field of aeronautical English assessment as a practitioner. I would also like to stress that I do not work for a company or for profit but for the Brazilian government. Receiving a master’s degree did not raise my income in any way. However, I acknowledge the influence of my profession on my research, especially from a practical point of view. Being a professional in this field contributed to my desire to follow a
more straight-forward, to-the point research approach (Dörnyei, 2007). In response, I decided to expend more energy on the more practical aspects of the research than on discussing the philosophical principles underpinning the project. This pragmatic perspective has had a major influence on how I planned this research (the questions I asked, the methods I used), as well as on how I interpreted and framed the results.

Additionally, although I am not a pilot, working as a public servant has given me privileged access to participants. Since ANAC has access to test takers’ contact information, it was easy for me to email pilots an invitation to participate in the study. However, for ethics-related reasons and to comply with Carleton University’s guidelines, I sent the invitation email from my Carleton account, not ANAC’s. Since the vast majority of pilots do not know me by name, in all probability, they did not know I work for ANAC. I would not want them to feel obliged to participate in my study.

Finally, I identify as a novice researcher. I am part of the ICAEA Research Group, of GEIA and of the International Language Testing Association (ILTA) Language Assessment in Aviation Special Interest Group (LAASIG). Being part of these communities of practice motivates me to work and contribute to the improvement of the quality of aeronautical English tests.

To conclude, I acknowledge that there are probably other aspects of my stance of which I am not aware that might have affected my research. However, I trust that this positionality statement gives readers an opportunity to consider how my personal values and experiences may have shaped and impacted the research presented in this dissertation from its beginning.
References


Garcia, A. C. M. (2015). 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. Master’s dissertation, Lancaster University.


https://commons.erau.edu/cgi/viewcontent.cgi?article=1021&context=icaea-workshop


ICAO (2022b, October). *ICAO language proficiency requirements: Achieving standardization*. Presentation by Catalin Popa [Video]. Skytalks. ICAO TV.

https://www.icao.tv/skytalks/season:1/videos/a41-skytalk-icao-language-proficiency-requirements-achieving-standardization


Chapter 2: Manuscript 1 – Investigating the construct of a listening test to assess pilots’ comprehension: A step-by-step project for test developers

The manuscript in this chapter consists of a conference proceedings article that has been published in the journal *Applied Linguistics Papers* published by the University of Warsaw. The guest editors were Prof. Anna Borowska, from the University of Warsaw, and Dr. Patricia Tosqui-Lucks, from the Airspace Control Institute (ICEA), in Brazil. This paper explains steps that may be followed in order to define the construct of a listening test in the aeronautical English context. It presents a framework that may be useful in a construct definition project for the development of a listening test for pilots and ATCOs while also providing insights for developers of other language assessments for professional purposes (LAPP).

The following corrections/suggestions of improvement should be acknowledged:

<table>
<thead>
<tr>
<th>Article page number/ line</th>
<th>From</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>15/12</td>
<td>criterion 3</td>
<td>Criterion 3</td>
</tr>
<tr>
<td>16/4</td>
<td>To conclude this introduction, some questions are suggested.</td>
<td>After this introduction, some questions that test developers may address are suggested.</td>
</tr>
<tr>
<td>16/8</td>
<td>The appendices include…</td>
<td>The project includes…</td>
</tr>
<tr>
<td>17/footnote 2</td>
<td><a href="https://www4.icao.int/aelts/Home/RecognizedTest">https://www4.icao.int/aelts/Home/RecognizedTest</a></td>
<td><a href="https://www4.icao.int/aelts/Home/RecognizedTests">https://www4.icao.int/aelts/Home/RecognizedTests</a></td>
</tr>
<tr>
<td>20/13</td>
<td>Stated and theoretical construct</td>
<td>Stated and theoretical constructs</td>
</tr>
<tr>
<td>-------</td>
<td>----------------------------------</td>
<td>-----------------------------------</td>
</tr>
<tr>
<td>(Table 2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22/17</td>
<td>STDQ 1 to 8</td>
<td>STDQ 1 to 5</td>
</tr>
<tr>
<td>(Table 3’s caption)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23/25</td>
<td>These two characteristics of LSP testing, may bring about</td>
<td>These two characteristics of LSP testing may bring about (no comma)</td>
</tr>
</tbody>
</table>

Thus, *validation* and *validity*

| 24/16 | As proposed in the provisional questions included in the appendices | As proposed in the provisional questions included in the appendices |

<table>
<thead>
<tr>
<th>29/12</th>
<th>Research the construct of aeronautical English tests</th>
<th>Research on the construct of aeronautical English tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Table 8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32/20</td>
<td>… on the proposed construction</td>
<td>… on the proposed construct</td>
</tr>
<tr>
<td>(Table 10)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Investigating the Construct of a Listening Test to Assess Pilots’ Comprehension: a Step-by-Step Project for Test Developers

Angela Carolina de Moraes GARCIA
Carleton University
E-mail: angelagarcia@cmail.carleton.ca

Abstract: Pilots and air traffic controllers must demonstrate their ability to listen and speak the language used in radiotelephony communications demonstrated by completing a language test. In this context, it is crucial to assess both interactive listening, when listening occurs together with speaking, and listening in isolation, when there is no speaking or interaction. The purpose of assessing listening in isolation is to reduce the influence of skills that are not relevant to the construct, that is, to minimize construct irrelevant variance (S. Messick 1989). This article describes a project that can be followed by test developers to address the initial step in the development of a test to assess pilots’ listening in isolation: the construct definition. The project is framed within an interactionalist perspective wherein a test construct is defined based on a combination of the abilities that those taking the test should have and the tasks that they should be able to perform (L. Bachman 2007). It is also informed by the work of L. Bachman/A. Palmer (2010) and the framework proposed by U. Knock/S. Macqueen (2020) for the development of language assessments for professional purposes. The project outlined in this article may also be of interest to test developers who wish to investigate different constructs of aeronautical English tests, as well as those involved in the development of other types of language assessments for professional purposes.

Keywords: language testing, language assessment for professional purposes, ICAO language proficiency requirements, listening comprehension, construct definition

Introduction

This paper illustrates one possible way to define the construct of a language test in the context of assessing languages for specific purposes (LSP), or more specifically, in the context of language assessment for professional purposes (LAPP). U. Knock/S. Macqueen (2020: 3) define LAPP as “any assessment process, carried out by and for invested parties, which is used to determine a person’s ability to understand and/or use the language of a professionally-oriented domain to a specified or necessary level.” Although the project outlined in this paper aims to illustrate the steps that could be taken to investigate the construct of a listening test in the context of aeronautical English testing, it may also be of interest to test developers who wish to investigate different constructs of aeronautical English tests, as well as those involved in the development of other types of language assessments for professional purposes. This paper also intends to bridge the gap between theory and practice by outlining practical steps that are informed by the latest research and language assessment theory.
The term aeronautical English refers to the English used by pilots and air traffic controllers (ATCOs) in their radiotelephony (RT) communications (P. Tosqui-Lucks/A. Silva 2020). According to the International Civil Aviation Organization (ICAO)’s language proficiency requirements (LPRs), pilots and ATCOs’ ability to speak and understand the English used in RT communications must be assessed. ICAO developed a rating scale to be used in such assessments, the ICAO rating scale, which includes six areas of evaluation: pronunciation, structure, vocabulary, fluency, comprehension, and interactions. Moreover, Criterion 3 of the ICAO test design guidelines, developed by the International Civil Aviation English Association (ICAEA), states that “test instruments need to contain tasks dedicated to assessing listening comprehension, separate from tasks designed to assess speaking performance” (ICAEA n.d.a). The proposed project aims to assess listening in isolation, as required by criterion 3. This type of assessment is necessary because it allows listening to be assessed by itself, without much interference from other skills. A central principle in language testing is to minimize construct irrelevant variance by reducing the influence of skills that are not relevant to the construct (S. Messick 1989). Although it is necessary to assess pilots and air traffic controllers’ listening comprehension in isolation, this should be done in addition to and not to the exclusion of the assessment of interactive listening during the speaking test, as, most of the time, pilots and ATCOs need to listen and to interact with each other simultaneously.

In order to develop a listening test to assess pilots’ listening in isolation, it is necessary to have a clear definition of the construct to be assessed. An important consideration is that listening happens inside our minds, so it cannot be assessed directly, as speaking and writing (J. Field 2019). L. Harding (2015: 123) argues that, in language testing, listening is “still a very under-represented skill”. In the context of aeronautical English assessment, a fairly new field of research, this problem is even more prominent. The process for creating a high stakes test, such as a listening test for pilots and ATCOs, is complex, time-consuming and requires a great effort from the needs analysis to the operationalization of the test, including the definition of the construct, the writing of test specifications, the development of tasks, the trial of test items, the development of the scoring criteria, and the validation of the test. The goal of the proposed project is to address the initial stage of the development of a test to assess pilots’ listening comprehension in isolation: the construct definition.

Thus, the proposed overarching question to be addressed in the proposed project is:  
What should be the construct of an aeronautical English listening test for pilots?

A similar question could be formulated for ATCOs and the project described here could be adapted for ATCOs rather than pilots. However, in keeping with the ICAO test design guidelines that “separate test instruments need to be designed for pilots and air traffic controllers” (Criterion 2, ICAEA n.d.b), different listening tests should be developed for pilots and ATCOs. In order to address this question, a needs analysis should be conducted. A needs analysis is an investigation of the target language use (TLU) domain (i.e., “a specific setting outside the test itself that requires the test taker to perform language use tasks” [L. F. Bachman/ A.S. Palmer 2010: 60]), and is an essential step in the development of an assessment of LSP, especially in the development of LAPPs. According to U. Knoch/ S. Macqueen (2020: 83), a careful needs analysis helps to “increase the
trustworthiness of an assessment instrument”. In conducting this needs analysis, “the contributions of policy, practices, selected social theories, empirical research, and multiple stakeholders” (A. Monteiro/ J. Fox 2022: 170) should be taken into consideration.

To conclude this introduction, some questions are suggested. In the next section, the suggested methodology is explained, including an explanation about the theoretical foundation that could underpin such project. The conclusion includes suggestions for future steps to be taken, some limitations of the proposed project, and some final considerations. The appendices include provisional interview, questionnaire and focus group questions that could be used for the project.

1. Specific questions that test developers may address

As A. Monteiro/ J. Fox (2022: 194) argue, “in multicultural professional contexts in which participants use ELF [English as a Lingua Franca] alongside workplace-specific terminology, such as international radiotelephony communications in aviation, (...) test contexts and constructs should be defined based on characteristics of the TLU domain anchored in the perspectives and accounts of domain stakeholders”. With this objective in mind, some specific test development questions were formulated as well as a methodology to address them. The suggested questions are listed in Table 1 and have been formulated within an interactionalist perspective wherein a test construct is defined based on a combination of the abilities that those taking the test should have and the tasks that they should be able to perform (L. Bachman 2007). As L. Bachman (2007: 42) explains, the interactionalist perspective “views the construct we assess not as an attribute of either the individual language users or of the context, but as jointly co-constructed and residing in the interactions that constitute language use”. Defining the construct based on the interaction between both abilities and tasks has not only been recommended (e.g., C. Chapelle 1998; M. Chalhoub-Deville 2003), but it has been argued to be the most appropriate approach in the case of listening tests where the listening performance is a result of the underlying knowledge and ability, the situational factors, and the interaction between them (G. Buck 2001). This approach is especially useful in the context of LAPPs because of the importance of the professional domain to the language assessment. As U. Knoch/ S. Macqueen (2020: 63) point out, “the goal of LAPP is to extrapolate an ability classification based on a brief performance sample in a relatively contrived context to other contexts of use in which professional knowledge is central”. Thus, I suggest following what L. Bachman (2007) calls a moderate interactionalist approach to construct definition: the “an ability – in language user – in context” perspective suggested by M. Chalhoub-Deville (2003). This approach highlights the importance of considering context when investigating test constructs because, according to M. Chalhoub-Deville (2003: 369), “individual ability and contextual facets interact in ways that change them both”.

Applied Linguistics Papers: www.alp.uw.edu.pl
Investigating the Construct…

Specific test development questions (STDQ)

<table>
<thead>
<tr>
<th></th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>What are the available resources to develop and administer a test to assess pilots’ listening comprehension in isolation, and what are the expectations of the organization which is developing the test or to which the test is being developed?</td>
</tr>
<tr>
<td>2</td>
<td>According to the ICAO regulations and guidelines, what types of skills, knowledge and processes should be assessed in a test that aims to assess pilots’ listening in isolation, and what are the TLU domain tasks and their characteristics?</td>
</tr>
<tr>
<td>3</td>
<td>According to the academic literature and research, what types of skills, knowledge and processes are needed in the TLU domain and thus should be assessed in a test that aims to assess pilots’ listening in isolation, and what are the characteristics of the TLU domain tasks?</td>
</tr>
<tr>
<td>4</td>
<td>What types of skills, knowledge, and processes are assessed in a recognized listening test for pilots, and what TLU domain tasks are represented in it?</td>
</tr>
<tr>
<td>5</td>
<td>According to key stakeholders (e.g., pilots and/or ATCOs, raters, researchers), what types of skills, knowledge and processes are needed in the TLU domain and should be assessed in a test that aims to assess pilots’ listening in isolation, and what are the characteristics of the TLU domain tasks?</td>
</tr>
<tr>
<td>6</td>
<td>Based on the accounts of aeronautical English researchers, how could the construct defined in the draft <em>design statement</em> be refined?</td>
</tr>
</tbody>
</table>

Table 1. Specific test development questions (STDQ).

The goal of STDQ 1 is to find out what resources are available and to learn about the needs and expectations of the organization that is requiring the development of the test. STDQ 2 aims to investigate the construct according to the ICAO regulations and guidelines (e.g., ICAO 2010; ICAO 2020), including the ICAO rating scale, and the ICAO test design guidelines, developed by ICAEA (ICAEA n.d.c)\(^1\). I believe a detailed inspection of the policy and the guidelines is a good starting point for the development of a listening test. As C. Moder/ G. Halleck (2021: 82) suggest, one of the first steps to be taken to develop a LSP test is “to consult relevant information provided by government agencies and by professional groups charged with the language training of the professionals in the target domain”. After that, I suggest carrying out secondary research to investigate how academic research and literature can contribute to the definition of the construct of a listening test for pilots, by conducting a thorough review of the literature (STDQ 3). STDQ 4 aims to investigate the construct of an existing listening test for pilots or ATCOs. At the time of writing this paper, the only test endorsed by ICAO\(^2\) was the ELPAC for ATCOs. The ELPAC, which stands for *English Language Proficiency for Aeronautical Communication*, was developed by the European Organisation for the Safety of Air Navigation (EUROCONTROL), in partnership with other institutions. Looking into the construct of an existing listening test may valuably inform the decisions to be made in relation to what

\(^1\) At the time this paper was written, a group of experts which ICAO invited were revising the ICAO test design guidelines in order for ICAO to have it published as a handbook.

\(^2\) See [https://www4.icao.int/aelts/Home/RecognizedTest](https://www4.icao.int/aelts/Home/RecognizedTest)
construct should be measured in a test to be developed. The goal of STDQ 5 is to investigate the perceptions of key stakeholders such as pilots and/or ATCOs on what the construct of a listening test should be, similarly to what A. García/J. Fox (2020) have done. It is pivotal to investigate the perceptions of domain experts in relation to the construct to be assessed as “language and communication may mean one thing to linguistically oriented professionals and another to gatekeepers within a professional community” (S. Jacoby/T. McNamara 1999: 236). Finally, STDQ 6 aims to investigate the perceptions of aeronautical English researchers about the construct that should have been defined based on the data gathered in the study. The purpose is to have the researchers evaluate the definition of the construct so it can be improved. If access to researchers is difficult, you may consider inviting other key stakeholders to evaluate the defined construct (e.g., aeronautical English test raters, teachers, etc.). Figure 1 shows an overview of the proposed project, including the specific test development questions each strand addresses.
2. Proposed methodology

In the proposed project, I follow a pragmatic approach, which advocates the use of “a straightforward, ‘to-the-point’ research methodology that offers a great deal of practical help” (Z. Dornyei 2007: 18). However, good research must follow a sound theoretical framework, as well as a strong methodological approach. Therefore, I recommend using as the conceptual foundation for the proposed project the socio-cognitive theory developed by C. Weir (2005). The socio-cognitive approach to testing, as explained by M. Milanovic/C. Weir (2013: x), “seeks to take account of both the aspects of cognition, related to the mental processes the individual needs to engage in order to address a task, and the features of language use in context that affect the ways in which a task is addressed”. I also suggest following the recommendations to develop language assessments for professional purposes given by U. Knoch/ S. Macqueen (2020). Other useful works which may underpin the project include L. Bachman/ A. Palmer (2010) and D. Douglas (2000). In relation to the methodological approach, needs analyses usually draw on multiple sources and methods (U. Knoch/ S. Macqueen 2020). Therefore, I recommend conducting a multiple methods study, in which the strengths of both qualitative and quantitative approaches can be combined in a complementary way (Z. Dornyei 2007), as gathering both qualitative and quantitative data helps to develop a better understanding of the research problem than using only one approach (J. Creswell 2015). U. Knoch / S. Macqueen (2020: 96) explain that “drawing on multiple data sources to triangulate the results increases the credibility of the conclusions drawn based on the needs analysis”. Thus, the proposed project uses three types of triangulation, as explained by J. Brown / T. Rodgers (2002): data triangulation, by using multiple sources of information (data from policy, literature, and different stakeholders – pilots/ATCOs, test developers, researchers); theory triangulation, by using multiple theoretical frameworks (D. Douglas 2000, M. Chalhoub-Deville 2003, L. Bachman/ A. Palmer 2010, U. Knoch/ S. Macqueen 2020), and; methodological triangulation, by using different procedures to collect data (literature review, interviews, focus group, questionnaires, document analysis).

2.1 The theoretical framework and its relationship with the proposed study

U. Knoch/ S. Macqueen (2020) propose a socially-oriented theory of construct for LAPP, a model of needs analysis, procedures for turning the results of a needs analysis into a test blueprint and specifications, as well as a framework to be used for validation of LAPP. For a better understanding of this proposed study, it is important to, firstly, define the four dimensions of construct proposed by U. Knoch/ S. Macqueen (2020): the stated construct, the operationalized construct, the theoretical construct, and the perceived construct.

- The **stated construct** refers to what is publicly claimed to be assessed on a test. Information about the stated construct can be found on the test’s website (description of the test, sample test), on the policy, the rating scale, etc.;
- The **operationalized construct** refers to what is really being assessed during the actual test. For researchers to gather information about the operationalized construct, they would have to have access to the actual performances of test takers (e.g., their responses) or to their behaviour during the test, in order to investigate what they are, for example, thinking or doing during the test;
Another dimension of construct, the *theoretical construct*, is unobservable and refers to the theory on which the assessment is based. U. Knoch/ S. Macqueen (2020: 40) explain that “in LAPP, this is typically a language proficiency, skill or ability that is assumed to underlie communication in an actual, specific world of work”. The theoretical construct might be explicitly stated or not. Whatever the case may be, information about it has likely guided the development and design of the test, including the test procedures;

Lastly, the *perceived construct* refers to how participants in the testing process (test takers, raters, policy makers, teachers, etc.) understand the construct (what they believe the test is testing).

Table 2 shows the spheres of construct related to the STDQ 2 to 5:

<table>
<thead>
<tr>
<th>STDQ</th>
<th>Sphere of construct</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td><em>Stated and theoretical construct</em></td>
</tr>
<tr>
<td>3</td>
<td><em>Theoretical construct</em></td>
</tr>
<tr>
<td>4</td>
<td><em>Stated and perceived constructs</em></td>
</tr>
<tr>
<td>5</td>
<td><em>Perceived construct</em></td>
</tr>
</tbody>
</table>

*Table 2. Sphere of construct related to STDQ 2 to 5 (U. Knoch/ S. Macqueen 2020).*

Thus, STDQ 2 addresses the stated construct in the ICAO regulations and guidelines, and also the theoretical construct explained especially in the ICAO test design guidelines. STDQ 3 aims to investigate the theory on which the listening test for pilots should be based. STDQ 4 investigates how the construct of a recognized test is stated on their website and publicly available documents, and how it is perceived by its test developer. Finally, STDQ 5 aims to investigate how key stakeholders (pilots and/or ATCOs) perceive the construct to be assessed. The sphere of operationalized construct is not included because it is only possible to start looking into it after the test starts being operationalized.

Secondly, U. Knoch/ S. Macqueen (2020) proposed a cycle of development of LAPP. As shown in Figure 1, the stages for the development of LAPP start with the needs analysis and end with the operational use of the test. However, there needs to be regular review, and the work at a certain stage may require that test developers go back to a previous stage. The proposed project addresses the first stage of the LAPP test development cycle, the needs analysis, and part of the second stage, the development of the design statement, as shown by the blue rectangle in Figure 2.
Thirdly, it is worth considering U. Knoch/ S. Macqueen’s (2020) approach to needs analysis. These researchers talk about five areas a needs analysis for developing a LAPP may address: the domain analysis, the means analysis, the policy analysis, the test requirement analysis, and the test taker analysis. The domain analysis is a key component in a needs analysis. It requires an empirical analysis of the TLU communication tasks, of the language used to complete these tasks, as well as of the real-world interaction between test takers and tasks. The means analysis is an analysis of the available resources for test development, administration and validation. The policy analysis is an analysis of the regulations that are relevant to the LAPP. The test requirement analysis is an analysis of the requirements for the test, such as test purpose, information about the needs of different stakeholders, and about score reporting and score uses. Finally, the test taker analysis is an analysis of the language proficiency of test-takers and of the difficulties they may encounter when communicating in real life, as well as an analysis of the test takers’ perceptions of the test, their needs and motivations, and the impact that the test might have on teaching. Table 3 shows the areas of needs analysis addressed by each specific test development question. As U. Knoch/ S. Macqueen (2020) do not mention a kind of needs analysis that aims to investigate the construct by looking at other tests, I have named this kind of analysis external tests analysis. U. Knoch/ S. Macqueen (2020) explain that these types of needs analyses do not necessarily need to follow a linear sequence or come before all
stages of assessment design. They follow a cyclical procedure. We can understand them as independent activities that overlap each other.

<table>
<thead>
<tr>
<th>STDQ</th>
<th>Areas of needs analysis addressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Means analysis</td>
</tr>
<tr>
<td></td>
<td>Test requirement analysis</td>
</tr>
<tr>
<td>2</td>
<td>Policy analysis</td>
</tr>
<tr>
<td></td>
<td>Test requirement analysis</td>
</tr>
<tr>
<td></td>
<td>Domain analysis</td>
</tr>
<tr>
<td>3</td>
<td>Policy analysis</td>
</tr>
<tr>
<td></td>
<td>Test requirement analysis</td>
</tr>
<tr>
<td></td>
<td>Domain analysis</td>
</tr>
<tr>
<td></td>
<td>Test taker analysis</td>
</tr>
<tr>
<td>4</td>
<td>External test analysis</td>
</tr>
<tr>
<td>5</td>
<td>Domain analysis</td>
</tr>
<tr>
<td></td>
<td>Test requirement analysis</td>
</tr>
<tr>
<td></td>
<td>Test taker analysis</td>
</tr>
</tbody>
</table>

Table 3. The areas of needs analysis for STDQ 1 to 8.

As U. Knoch/ S. Macqueen (2020: 108) point out, “needs analyses generally result in an array of information which is often difficult to consolidate”. They suggest, similarly to what L. Bachman/ A. Palmer (2010) had suggested, to start with the development of a draft design statement based on the information gathered during the needs analysis. The design statement should contain information to guide the next stages of test development, including a description of the purpose of the test, a description of the test-takers, a description of the TLU, and the definition of the construct to be measured. The information contained in this document should also serve as evidence (or backing) for the warrants in the validation framework. L. Bachman/ A. Palmer (2010) give a more detailed structure for a design statement. They suggest, for example, that a design statement should also include a list of “tasks selected as a basis for developing assessment tasks,” as well as a “description of the characteristics of the TLU tasks that have been selected as a basis for assessment tasks” (L. Bachman/ A. Palmer 2010: 270). U. Knoch/ S. Macqueen (2020) suggest that this list of tasks, which they call a table of TLU tasks, should be created after the design statement is produced. However, this list (or table) can be incorporated in the design statement, as recommended by L. Bachman/ A. Palmer (2010). Thus, Table 4 shows the information that the design statement should include:

---

Although other authors (e.g., J. Alderson et al. 1995, R. Green 2017) have recommended that the definition of the construct should be included in the test specifications document, the project explained in this paper follows L. Bachman and A. Palmer’s (2010) guidelines. The construct should be defined in the design statement. Future blueprint should include the assessment specifications and the task specifications (including, the construct for each task type).
Structure of Design Statement

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Purpose of the test</td>
</tr>
<tr>
<td>2</td>
<td>Description of the test takers</td>
</tr>
<tr>
<td>3</td>
<td>Definition of the construct</td>
</tr>
<tr>
<td>3.1</td>
<td>Description of the TLU domain</td>
</tr>
<tr>
<td>3.2</td>
<td>Types of skills, knowledges and processes needed in the TLU domain.</td>
</tr>
<tr>
<td>3.3</td>
<td>List of TLU domain tasks selected as a basis for developing assessment tasks</td>
</tr>
<tr>
<td>3.4</td>
<td>Description of the characteristics of the TLU domain tasks</td>
</tr>
</tbody>
</table>

Table 4. Information that might be included in the Design Statement.

The detailed explanation given by L. Bachman/ A. Palmer (2010) about the procedures to be followed to develop a design statement should also be very helpful, as well as the projects they have shared4. Although L. Bachman/ A. Palmer (2010) list the sections on the description of the TLU domain, on the tasks, and on the characteristics of TLU tasks separately, I have joined them under the same section because according to the interactionalist approach on construct definition, the construct should be defined based on both traits and context. For section 3.2 of the design statement, I suggest following D. Douglas’s (2000) framework of components of specific language ability. For section 3.4, I recommend adapting L. Bachman/ A. Palmer’s (2010) framework of test task characteristics, because, as they indicate, their framework may be “useful for describing both TLU tasks and test tasks” (L. Bachman/ A. Palmer 1996: 57).

When designing the interview and questionnaire questions to be used, two issues related to the assessment of LSPs should be taken into consideration. They are specificity of content and inseparability. As discussed by D. Douglas (2001), these two characteristics of LSP testing, may bring about some theoretical and practical problems5. First, the issue of specificity deals with the problem of how specific test tasks should be. For example, should there be a test for all pilots, or a test for airplane pilots and another one for helicopter pilots? If a test was designed for airplane pilots only, there should still be issues because airplane pilots have different flying experiences, they fly different airplanes, and so on. So, the question here again would be how specific should the test be? Second, the problem of inseparability handles the understanding that general purposes language tests should not include the assessment of background knowledge, because that would be considered irrelevant for the construct. However, as D. Douglas (2001) argues, in LSP testing, it might not be possible to separate language knowledge from specific purpose background knowledge. For this reason, he advocates that “we must, in testing language for specific purposes, define specific purpose language ability as comprising both language

---

4 Available at http://www.oup.com/LAIP.
5 D. Douglas (2001) also discusses a third problem in assessing LSP: authenticity. In the project explained in this paper, the issue of authenticity does not need to be addressed because test tasks will not yet be developed; the proposed project only gets to the point of listing the TLU tasks and describing their characteristics.
knowledge and background knowledge” (D. Douglas 2001: 50). Thus, we need to ask ourselves to what extent should the test assess professional knowledge?

Lastly, although this paper proposes a project to be followed at the beginning of the test development cycle, issues related to validity should be discussed. As L. Bachman/ A. Palmer (2010) argue, the main purpose of language assessments is to collect information for making decisions that will ideally lead to beneficial consequences for stakeholders. Test developers need to be accountable to stakeholders because many tests, including the aeronautical English tests, are high stakes. L. Bachman (2015: 7) explains that high stake tests “have major, life changing consequences for stakeholders, and decision errors (false positive/negatives) are difficult to reverse”. Being accountable to stakeholders means that test developers need to be able to justify the use they make of an assessment. In other words, they need to show to their stakeholders that the intended uses of their assessment are justified (L. Bachman/ A. Palmer 2010). Stakeholders are all those involved in or affected by the assessment. In the aeronautical English context, stakeholders can be, to name a few, test takers, test developers, raters, regulators, aeronautical English teachers, airlines, passengers, and society in general. Thus, validation and validity are central concepts in language assessment. Validation may be defined as “the ongoing process of justifying particular interpretations and uses of test results” (C. Chapelle 1998: 33). Validity is an abstract theoretical term which was defined by S. Messick (1989: 13) as “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment”. Thus, validation is an activity through which validity is investigated. To guide this process of validation, a conceptual framework is necessary. M. Kane (2002, 2006) presents an argument-based approach to validity. As C. Chapelle et al. (2008: 6) explain, “arguments are used to build a case for a particular conclusion by constructing a chain of reasoning in which the relevance and accuracy of observations and assertions must be established and the links between them need to be justified”. An argument-based approach consists of a systematic way to investigate the validity of the interpretations and uses of scores (we should not claim that a test is valid, but we could claim that the interpretations and uses of test scores are valid). As U. Knoch/ S. Macqueen (2020) point out:

Argument-based approaches to validation rely on specifying a series of inferences, warrants and assumptions associated with score interpretations and uses. Inferences connect a series of claims we make. Underlying each inference, warrants and assumptions are formulated which need to be supported by evidence (also referred to as backing) so we can argue that the inference is supported (U. Knoch/ S. Macqueen 2020: 139).

U. Knoch/ S. Macqueen (2020) propose a validation framework to be used in LAPP. Their validation framework is underpinned by M. Kane’s (2002, 2006) work in educational measurement and by language assessment theories (C. Chapelle et al. 2008, L. Bachman/ A. Palmer 2010, U. Knoch/ C. Chapelle 2018). Their framework “is useful in that it provides an overarching framework which is connected by a series of inferences leading from the domain, the assessment materials to the test consequences” (U. Knoch/ S. Macqueen 2020: 164). U. Knoch/ S. Macqueen’s (2020) framework includes infor-
mation on both the competences and the context for score interpretation and use, in accordance with the interactionalist perspective to construct definition. The structure of U. Knoch/ S. Macqueen’s (2020) validation framework includes seven inferences: domain description, evaluation, generalization, explanation, extrapolation, decisions, and consequences (U. Knoch/ S. Macqueen 2020: 139). Their proposed framework is not fixed; it should be adapted to each particular testing context. The domain description inference (claim, warrants and assumptions) proposed by U. Knoch/ S. Macqueen (2020) highlights the importance of selecting, designing, and delivering assessment tasks that reflect the characteristics of the TLU domain. The assumptions specify that test tasks and assessment conditions should mirror the TLU domain and should sufficiently represent it. They also emphasize that professional knowledge should be included in the test “to the extent that it is required by the policy environment and test purpose” (U. Knoch/ S. Macqueen 2020: 143).

Although U. Knoch / S. Macqueen (2020) only discuss validity in Chapter 5 of their six-chapter book, the issues of validity and validation should be considered from the initial stages of test design and not only after the test is in use (G. Fulcher/ F. Davidson 2007). Thus, when designing the instruments to be used in this proposed study, validation issues should be taken into consideration. The domain description assumptions of U. Knoch/ S. Macqueen’s (2020) validation framework should inform the development of the questions for the interview and questionnaires (as proposed in the provisional questions included in the appendices). Thus, the evidence to be collected during the execution of this study may help to support the domain description assumptions.

2.2 Methods

The design of the proposed project consists of a multistep mixed-methods study in which qualitative data is dominant. Step 1 aims to gather useful information about the available resources and expectations of the organization developing the test or to which the test is being developed. As J. McDonough (1984) and J. Swales (1988) suggested, this needs to be taken into consideration at the beginning of the needs analysis because possible constraints may be identified from the initial stages of test development. The goal of Step 2 is to investigate the relevant skills, knowledges and processes that should be assessed in a test that aims to assess pilots’ listening in isolation as stated in the ICAO policy and as discussed in the relevant literature, as well as to investigate the TLU tasks and their characteristics. The results of the data analyses of Step 2 should inform the instruments to be used in Step 3, which aims to investigate the skills, knowledges and processes assessed by a recognized test of pilots or air traffic controllers. Step 3 also aims to investigate the TLU domain tasks that are represented in the recognized test. The purpose of Step 4 is to investigate the perceptions of key stakeholders (pilots, air traffic controllers, raters, and/or researchers) on the listening construct of the test. In A. Garcia/ J. Fox (2020), pilots responded to a questionnaire on the listening construct of a test to assess pilots’ listening comprehension (Phase A), and then aeronautical English experts, including raters, pilots, air traffic controllers and researchers, were interviewed (Phase B). After Step 4, you should have enough information to produce the draft design statement. Then, in Step 5, aviation English researchers or another group of experts should evaluate the design statement so that it can be improved.
Table 5 shows the types of data and their sources for each STDQ. The data gathered in Step 2 can be combined in a grid and these data should inform Step 3, and then the data collected in Step 3 can feed the grid to inform Step 4. The grid with the results of Steps 2, 3 and 4 should be used to elaborate the design statement. Data collected and analysed in Step 5 should be used to refine the design statement.

<table>
<thead>
<tr>
<th>STDQ</th>
<th>Type of data</th>
<th>Sources of data or data collection methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>QUAL</td>
<td>Interview with test developer or institution requesting the test to be developed</td>
</tr>
<tr>
<td>2</td>
<td>QUAL</td>
<td>ICAO policy and ICAO test design guidelines developed by ICAEA</td>
</tr>
<tr>
<td>3</td>
<td>QUAL</td>
<td>Academic literature and research</td>
</tr>
<tr>
<td>4</td>
<td>QUAL</td>
<td>ELPAC’s website, ELPAC published documents, and e-mail interview with ELPAC test developer</td>
</tr>
<tr>
<td>5</td>
<td>quan, QUAL</td>
<td>Questionnaire with key stakeholders, Interviews with key stakeholders</td>
</tr>
<tr>
<td>6</td>
<td>quan, QUAL</td>
<td>Questionnaire with Aviation English Researchers, Focus group</td>
</tr>
</tbody>
</table>

Table 5. Type of data and sources of data per STDQ.

A. Monteiro/ J. Fox (2022) argue that

Test development and construct specification are strengthened when they are informed not only by theory and empirical research, but also by transdisciplinary stakeholders (e.g., pilots, ATCOs, test developers, trainers) whose expertise is rooted in varying lived experience of the construct as it plays out in actual practice. Improved construct specification leads to tests that are more aligned with the communicative needs of test takers and, as a result, have fewer unintended consequences (A. Monteiro/ J. Fox 2022: 165).

Therefore, the suggested participants in this proposed study are transdisciplinary stakeholders: pilots, ATCOs, researchers, test developers, English Language Experts (ELE) raters, with background in teaching English as a second language, Subject Matter Experts (SME) raters, who are experienced pilots or ATCOs, and aeronautical English researchers (see Table 6). Purposive sampling (L. Cohen et al. 2011) can be used in order to choose the participant to be interviewed in Step 3 (developer of a recognized test).
Investigating the Construct…

<table>
<thead>
<tr>
<th>STDQ</th>
<th>Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Test developer or test requester</td>
</tr>
<tr>
<td>4</td>
<td>Developer of a recognized test</td>
</tr>
</tbody>
</table>
| 5    | Phase A: airplane and helicopter pilots, ATCOs  
     | Phase B: Pilots, ATCOs, raters, researchers. In A. Garcia/ J. Fox (2020), 156 pilots answered a questionnaire in Phase A, and six experts were interviewed in Phase B: Four pilots (including one who holds a PhD in Linguistics and is an active researcher in the field of aeronautical English); one ATCO (all of whom have had experience working as ICAO test raters); one expert in aeronautical English who has worked as a test developer, rater trainer and test administrator and who was an active researcher in the field of aeronautical English |
| 6    | Aeronautical English researchers |

Table 6. Participants in the study per STDQ.

In Step 1, qualitative data should be collected through an interview with the test developer or test requester about the resources for the development of a listening test for pilots, about the possible conditions of test administration, and about the test developers’ needs. The interview can be conducted through email, as suggested by E. Dahlin (2021). The advantage of conducting the interview through email is that the participant has time to elaborate on the questions and give richer responses. Table 7 shows examples of questions that may be asked in this interview.

<table>
<thead>
<tr>
<th>Since the ICAO test design guidelines developed by ICAEA state that “test instruments need to contain tasks dedicated to assessing listening comprehension, separate from tasks designed to assess speaking performance”, please answer the following questions:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
</tbody>
</table>

Table 7. Examples of questions that may be asked in the interview with the test developer/requester.

Data gathered during this step may help to inform sections 1 and 2 of the design statement. Section 1 (“purpose of the test”) should also be based on the ICAO policy and ICAO test design guidelines developed by ICAEA. Studies asking demographic questions similar to A. Garcia (2017) can also be used to inform Section 2 (“description of the test takers”). Figure 3 shows the diagram of Step 1.
In Step 2, as seen in Figure 4, qualitative data from the policy and from the literature should be combined and used to inform both the instruments to be used in the next steps and the design statement.

In the literature review, I suggest looking into the following topics: language assessment in general, assessment of listening, some relevant social theories, LSP, and aeronautical English, having a closer look at other studies on the construct of aeronautical English tests. Table 8 shows examples of relevant literature that could be used in this review:
<table>
<thead>
<tr>
<th>Topic</th>
<th>Examples of relevant literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research the construct of aeronautical English tests</td>
<td>H. Kim 2013, A. Monteiro 2019, M. Park 2021, Silva 2022</td>
</tr>
</tbody>
</table>

Table 8. Literature review to be conducted in Step 2 (STDQ 3).

Figure 5 shows a diagram that might be followed in Step 3. In this step, I suggest starting by collecting data about the construct of the recognized test from information available on the test’s website and publicly available documents. Then, the data should be analysed. A table with the information of the test’s stated construct can be produced, and the draft interview questions can be refined. Next, the e-mail interview with the test developer can be conducted. After that, the data should be analysed and merged with the results of steps 1 and 2.

![Figure 5. Step 3 Diagram (STDQ 4).](image)

Table 9 shows examples of questions that may be asked in the interview with the developer of a recognized test. Questions 2, 5, 6, and 7 are related to the domain description assumptions 1, 2, 3, and 5, respectively, from the validity framework proposed by U. Knoch/ S. Macqueen (2020).
Please answer the following questions in as much detail as you can:

1. What skills, knowledges, processes, and strategies does the listening test assess?

2. Do you feel assessment tasks mirror those in the TLU domain? How?

3. What do you feel the test is assessing well? And what do you feel the test is not assessing well?

4. Do you feel there is significant construct irrelevance variance (something that is being assessed but should not be)?

5. Do you feel the test is not assessing something that it should be assessing (construct underrepresentation)? In other words, do you feel that the chosen assessment tasks sufficiently represent the TLU domain?

6. Do you feel that the assessment tasks elicit and are sufficiently representative of the types of skills, knowledges and processes needed in the TLU domain?

7. How does the test incorporate the domain-specific professional knowledge? To what extent is technical knowledge included in the test?

<table>
<thead>
<tr>
<th>Table 9. Examples of questions that may be asked in the interview with the recognized test developer.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 6 gives an example of a diagram that might be followed in Step 4. This diagram of an empirical mixed-method study was followed in a pilot study I conducted during my PhD program. (A. Garcia/ J. Fox 2020). This study applied a two-phase explanatory sequential design, as defined by J. Creswell (2015). The first phase was the collection and analysis of the quantitative data (questionnaire answered by 156 pilots), followed by the collection and analysis of the qualitative data (interviews with six key stakeholders). The qualitative data explains or expands on the results of the quantitative data. In this mentioned study, the questionnaire was hosted on Qualtrics, whereas the interviews were semi-structured and conducted through Skype. After Step 4, the results from steps 2, 3, and 4 can be combined to draft section 3 of the design statement: the construct definition of the test.</td>
</tr>
</tbody>
</table>
Step 5 may start with the presentation of the research project to the participants (either online or in person). In this presentation, the draft design statement detailing the test’s construct definition should be explained. Next, participants can answer a questionnaire about the draft design statement. Examples of questions that can be asked in this questionnaire can be seen in Table 10. Questions 6, 7, and 8 are related to the domain description assumptions 1, 3, and 5, respectively, from the validity framework proposed by U. Knoch/ S. Macqueen (2020).
To what extent do you agree or disagree with the following statements?

<table>
<thead>
<tr>
<th></th>
<th>( ) I strongly disagree with it</th>
<th>( ) I disagree with it</th>
<th>( ) I somewhat disagree with it</th>
<th>( ) I somewhat agree with it</th>
<th>( ) I agree with it</th>
<th>( ) I strongly agree with it</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The proposed definition of the construct of a test to assess pilots’ listening in isolation is complete and is not excluding any relevant skills, abilities and processes.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>The proposed definition of the construct of a test to assess pilots’ listening in isolation is not including the assessment of skills, abilities and processes that are irrelevant to the TLU domain.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>The list of TLU domain tasks selected as a basis for developing assessment tasks sufficiently represent the TLU domain.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>The list of TLU domain tasks selected as a basis for developing assessment tasks are sufficiently representative of the types of skills, knowledges and processes needed in the TLU domain.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>A test to be developed on the basis of the proposed construct definition will likely assess competencies which are important for pilots when they listen to air traffic control communications.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>The assessment tasks to be created based on the proposed construction definition will likely mirror those in real-life pilot/ATCO communications.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>The list of tasks is sufficiently representative of the types of skills, knowledge and processes that pilots need when listening to radiotelephony communications.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 10. Examples of questions that may be asked in the questionnaire.

The questionnaire data should be analysed, and the focus group interview questions refined. Then, the participants should discuss in a focus group the results of the questionnaire and the improvements to the design statement. Table 11 shows examples of questions that may be asked in the focus group.
We are going to discuss the following questions:

<table>
<thead>
<tr>
<th></th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>What is your opinion about the construct represented in the proposed document?</td>
</tr>
<tr>
<td>2</td>
<td>In your opinion, how can the proposed construct definition be improved?</td>
</tr>
<tr>
<td>3</td>
<td>In your opinion, does the Design Statement include the need to assess anything that you believe to be UNNECESSARY?</td>
</tr>
<tr>
<td>4</td>
<td>In your opinion, does the Design Statement NOT include the need to assess something that you believe to be NECESSARY?</td>
</tr>
<tr>
<td>5</td>
<td>Have you identified any technical or language mistake in the Design Statement?</td>
</tr>
<tr>
<td>6</td>
<td>How do you think the implementation of the test will impact aviation safety?</td>
</tr>
</tbody>
</table>

*Table 11. Examples of questions that may be asked in the focus group.*

The focus group interview should be recorded, and the recordings should be transcribed and analysed. With the results of Step 5, you can refine the draft design statement and finalize the test construct definition. Figure 7 shows the mixed-methods diagram for Step 5.
The qualitative data can be analysed through descriptive coding (J. Saldaña 2013) and the quantitative data generated through questionnaires can be analysed through descriptive statistical analyses. As previously mentioned, U. Knoch/ S. Macqueen’s (2020) validation framework informed the development of the suggested provisional questions for the interview, questionnaire, and focus group.

3. Next steps and conclusion

U. Knoch/ S. Macqueen (2020) suggest that after the table of domain tasks is ready, test developers should examine in detail the list of tasks and check whether they are useable, suitable, workable and how likely they will result in positive test preparation behaviours and effects. In order to scrutinize this information, some questions that test developers might ask are: How much language does this task require? How easy is it to accomplish this task? Does this task pose a threat to safety? Does this task need to be adapted? Does this task rely too much on background knowledge? Is this task too specific for use in a test? How may this task influence teaching and learning? Then, test developers would be
able to start to produce the blueprint, which should include the specifications for the test and for the tasks. The development of such document is “a central and crucial part of the test construction and evaluation process” (J. Alderson et al. 1995: 9). Other critical questions to be asked that will inform the test specification and blueprint document are related to the test procedures, including the scoring method. Test developers will also need to consider how many times the test takers will be able to listen to the test prompts and how that will work (if they are to be allowed to listen to a prompt more than once, or if they will have to show they need to listen to it again by asking for clarification, or if the recordings will be played twice anyway), as in real-life situations pilots are expected to ask for clarification. Also, test developers will need to decide on what accents to include. In relation to this, E. Wagner (2022) points out:

The issue of accent variety is an important consideration for L2 listening test developers because they need to consider what accent or variety to use on their tests. The obvious answer would seem to be to use the “standard” variety of the language where the test is taking place, but in an age of globalization and multiculturalism, choosing the most appropriate accent varieties to use on an L2 listening test can be challenging (E. Wagner 2022: 228).

In designing a test to assess Brazilian pilots, test developers could check, for example, the most frequent international routes that Brazilian pilots take in order to define the most frequent accents to be included in the test and its proportion. Once the draft of test specifications and blueprint document is written, test developers can create a sample test and write enough items to pilot the test with a representative number of test takers. Aeronautical English corpora can be very useful in the creation of test items (A. Pacheco et al. 2020).

After having given suggestions of future research, I need to acknowledge some limitations of the project explained in this paper. As most parts of the gathered data are qualitative, the results might be biased by the subjective interpretation of the researcher. Also, there are several other ways a needs analysis could be conducted. This is just one possible way of doing it. Furthermore, I recognize that it is not possible to fully address all research questions or to review all existing relevant literature. This reminds me of U. Knoch/ S. Macqueen’s (2020) reflection:

There is never a definite outcome from a needs analysis, rather researchers are required to decide, within a policy environment and certain fixed test requirements, how workplace communication can best be presented in a series of test tasks. There is therefore no point in searching for the ‘truth’ in the data collected, but, … , the researcher needs to create their current-best-shot at what a test should look like (U. Knoch/ S. Macqueen 2020: 108).

To conclude, needs analysis should not only happen in the initial stage of test development. It should be repeated to make sure the test continues to represent the TLU adequately (U. Knoch/ S. Macqueen 2020). Moreover, in spite of the different goals practitioners and language test researchers may have, a clear definition of the construct is extremely important, as L. Bachman (2007) thoughtfully points out:

Perhaps the most important distinction between the roles of language testing researcher and practitioner is that of purpose, or goal. The language testing researcher's
goal, I believe, is to better understand, inter alia, the psychological and contextual factors that affect performance on language assessments, the types of language use that language assessments elicit, the relationship between language use elicited in assessments and that created in real-life settings, and the relationship between the abilities engaged in language assessments and those engaged in real-life settings. I would argue that the goal of the language testing practitioner, on the other hand, is to design and develop language assessments that are useful for their intended purposes. In either role, I believe that it is essential that we clearly define what it is we want to measure or what we want to investigate (L. Bachman 2007: 66).

If, like me, you are both a practitioner and a language test researcher, I hope that you can reconcile both of our goals by conducting an academically sound study that aims to clearly define what a listening test for pilots should assess in order to inform your practice as test developers. R. Green (2017: 29) states that “defining the construct accurately and reliably is arguably one of the most important responsibilities of test designers”. I trust we can accomplish this goal responsibly.

References

Field, J. (2019), Rethinking the second language listening test – from theory to practice. Sheffield.
Garcia, A.C.M. / J. Fox (2020), Contexts and constructs: Implications for the testing of listening in pilots’ communication with air traffic controllers, (in:) “The ESPecialist” 41(4), 1–33.
ICAO (2010), Manual on the implementation of ICAO language proficiency requirements (Doc 9835) (2nd ed.). International Civil Aviation Organization.


Rost, M. (2016), *Teaching and researching listening* (3rd ed.). London.


Chapter 3: Manuscript 2 – The listening construct: Theories and implications for the assessment of pilots and ATCOs

The manuscript in this chapter consists of a book chapter published in an e-book entitled “Aviation English – A Global Perspective: Analysis, teaching, assessment” edited by Patricia Tosqui-Lucks and Juliana Santana. This e-book was published in 2022 as the Proceedings of the 8th GEIA Seminar, which was conducted online in November 2021. This chapter comprises a literature review on the nature of listening, the assessment of listening, and the specific features of the listening that takes place in aeronautical radiotelephony communications.

Alternatively, the chapter can be read by downloading the ebook through the following hyperlink:


The following corrections/suggestions of improvement should be acknowledged:

<table>
<thead>
<tr>
<th>Chapter page number/line</th>
<th>From</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>217/25</td>
<td>Acknowledgements</td>
<td>Acknowledgments</td>
</tr>
</tbody>
</table>
THE LISTENING CONSTRUCT: THEORIES AND IMPLICATIONS TO THE ASSESSMENT OF PILOTS AND ATCOS

Angela Carolina de Moraes Garcia
Carleton University, Ottawa, Ontario, Canada

ABSTRACT – Listening is complex. Assessing an invisible cognitive operation such as listening is challenging, especially in a high-stakes testing context. The purpose of this chapter is to discuss the main theories that have informed the definition of the listening construct in language testing (e.g., BUCK, 2001; FIELD, 2019) as well as to discuss some implications for the testing of pilots’ listening comprehension, as required by the International Civil Aviation Organisation (ICAO) policy. The main characteristics of the listening construct on a theoretical level are presented, followed by a discussion on issues related to its operationalization. It gives insights into what listening tests should be testing. Some features of the language used by pilots and ATCOs that may help their communication to be more effective are discussed, as well as those that may contribute to communication challenges. This discussion might be useful for listening test developers, especially those who are involved with the development of aeronautical English listening tests.

KEYWORDS: Assessment of listening; Construct definition; ICAO language proficiency requirements; Aeronautical English.

1. INTRODUCTION

When a language test needs to be developed, a basic question to be asked is: what is the test construct? In other words, test developers need to determine what they want to assess (BACHMAN, 2007). Understanding the construct is fundamental to test development. The term construct can be understood as a theoretical description of the relevant skills and abilities of a specific assessment context (FULCHER; DAVIDSON, 2007). CRONBACH & MEEHL (1955)’s definition of construct was “some postulated attribute of people, assumed to be reflected in test performance” (p. 283). Other definitions include: “a meaningful interpretation of observed behaviour” (CHAPELLE, 1998, p. 33), “the thing we are trying to measure” (BUCK, 2001, p. 1), “the theoretical entity that the test developers and test users intend the test to measure, the quality or qualities of the test takers we wish..."
to make inferences about” (DOUGLAS, 2010, p. 33), and “the trait (traits) or underlying ability that we intend to measure through assessment” (CHENG; FOX, 2017, p. 224). As Bachman (1990) explained, a clear definition of language abilities must be the basis of all language tests. However, defining a test construct may be a challenge (FOX, 2007), or even “a persistent problem” (BACHMAN, 2007, p. 41). No matter how difficult it may be, understanding the test construct is crucial to the process of language test development. As Chalhoub-Deville (2003, p. 369) argued, “the L2 construct is far reaching in terms of its impact on varied aspects of test design, validation, research, and theory formulation”. Thus, having a clear definition of the test construct is extremely important, especially in high-stakes testing contexts, such as the case of pilots/air traffic controllers (ATCOs) language proficiency testing.

The language proficiency requirements (LPRs) for pilots and ATCOs were introduced by the International Civil Aviation Organization (ICAO) in 2003 (ICAO, 2010). From March 2008 (later postponed to 2011), pilots and ATCOs should demonstrate their ability to speak and understand English (or the language normally used by the station on the ground) in order to be allowed to fly internationally (in the case of pilots) or control international flights (in the case of ATCOs). The ICAO LPRs include a set of holistic descriptors and a rating scale, which range from levels 1 to 6 in six different categories: pronunciation, structure, vocabulary, fluency, comprehension, and interactions. Pilots and ATCOs need to be awarded at least a level 4 in all six categories in order to receive a final level 4, the minimum required level for safe radiotelephony communication. Since the publication of the ICAO LPRs, some researchers have questioned the validity of the construct which underpins the ICAO policy and its rating scale (e.g., DOUGLAS, 2004, 2014; EMERY, 2014; KIM; ELDER, 2015; KIM; BILLINGTON, 2016; KNOCH, 2014; READ; KNOCH, 2009). They all expressed the importance of conducting more research to investigate the nature of the language used in pilots/ATCOs communications in order to inform construct definition. Douglas (2004, p. 250), for example, argued that “the combination of elements that must be incorporated into a construct definition of what is to be measured by the new tests is very complex, and it must be said, somewhat ambiguous in places”. Emery (2014) also considered the ICAO policy and rating scale to be problematic. He argued that “the ICAO guidance material … is of little practical use in the definition of the construct and the development of test specifications” (p. 206). Moreover, although ICAO (2010, p. 4-13) acknowledged that comprehension “represents half of the linguistic workload in spoken communications”, the fact that comprehension is, as mentioned, only one out of six skills in a rating scale which was developed to assess speaking may diminish the importance of the assessment of listening (GARCIA, 2015). Therefore, better understanding the listening construct in the context of pilots/ATCOs communications is crucial. It is important to note that comprehension is important for both pilots and ATCOs; however, it seems to be even more prominent for pilots, as ATCOs’ turns tend to be longer than pilots’ (CLARK et al., 1990).
As Hart (1998) argued, it is important to understand the history of the subject we are studying and learn how it has developed. Thus, the primary purpose of this chapter is to discuss the main theories that have informed the definition of the listening construct in language testing. I start by presenting the main characteristics of the listening construct on a theoretical level, followed by a discussion on issues related to its operationalization. Next, I discuss some implications for the testing of pilots’ listening comprehension. Throughout the chapter, I include some comments on construct validity (i.e., the validity of the inferences made from test scores) and on test validation, which can be understood as “the ongoing process of justifying particular interpretations and uses of test results” (CHAPELLE, 1998, p. 33). The chapter ends with a short conclusion.

It is important to emphasize that the comments on the assessment of pilots’ listening comprehension are related to the listening they perform in their communication with the ATCOs. In the workplace, pilots interact with a range of different professionals. Apart from ATCOs, they interact, for example, with flight attendants, flight dispatchers and aircraft maintenance personnel. Pilots also need to interact with each other, and, because of the increase in the number of flights with pilots that do not share the same first language, English has been increasingly used in the cockpit by pilots who are speakers of English as a second or foreign language. In this chapter, I refer to the language used by pilots and ATCOs in their communication as aeronautical English because, as Tosqui-Lucks and Prado (2020) explained, the term that has been traditionally used, aviation English, refers to a broader and more general context. According to them, “aviation English consists of the language used by all professionals in the aviation industry, such as mechanics, meteorologists, pilots and controllers” (p. 3).

It is also important to note that, although according to ICAO’s policy, pilots and ATCOs who are native speakers of English do not need to be formally evaluated, research has shown that being proficient in conversational English is not the same as being a proficient user of aeronautical English (TRIPPE, 2018, 2019). Estival et al. (2016, p. 199) claimed that “pilots who are native English speakers commit, in some cases, as many communication errors as English as a second language pilots”. As Read and Knoch (2009) argued, the ICAO LPRs have given “native-speaking aviation personnel no incentive to develop their communicative competence in ELF [English as a lingua franca] terms” (p. 21.7). For Douglas (2014), pilots and ATCOs who are native speakers of English should be assessed for their communication ability, and this assessment should include linguistic awareness and the “abilities to accommodate their use of English in the context of intercultural communication” (p. 2).

2. THE CONSTRUCT OF LISTENING

According to Buck (2001), defining the construct is a two-stage
Firstly, we define it on a theoretical or conceptual level, and then secondly, we operationalise that through the texts we select, and the tasks we ask our listeners to perform. The conceptualisation of our listening construct will be based on both our theoretical understanding of listening comprehension, and from our knowledge of the target-language use situation. The operationalisation of the construct will be based on our knowledge and experience as test developers. (BUCK, 2001, p. 94)

Then, the first step we should follow if we want to define a listening construct is to try to understand how the complex process of listening works. As Green (2017) pointed out, “it is essential that test developers spend quality time thinking about what the complete listening process involves before they embark on any task development work” (p. 1). Thus, in this section I start by presenting some characteristics of the listening skill on the conceptual level. Then, I explore issues related to the definition of the listening construct on the operational level. While discussing these topics, I include some considerations related to the assessment of pilots’ listening comprehension.

It is important to point out that, unfortunately, “very little is written in the language assessment literature on the specific constructs, or abilities, that underlie listening, on how to go about designing listening assessment procedures, on how to validate and evaluate listening tests” (ALDERSON; BACHMAN, 2001, p. x). Harding (2015) also highlighted that listening is “still a very under-represented skill” (p. 123).

2.1. What does listening consist of?

Language proficiency is traditionally understood as composed of four main language skills: writing, reading, speaking, and listening. Listening seems to be the skill that people spend most time using, about 45% of the total time of communication (FEYTEN, 1991, as cited by BUCK, 2001). Both reading and listening cannot be assessed directly, as they happen inside the mind of the listener (FIELD, 2019). However, “listening,” as Lynch and Mendelsohn (2010) pointed out, “is not merely an auditory version of reading” (p.180). We can no longer view listening as a passive or receptive skill because the successful listener plays an active role in understanding what was said. Anderson and Lynch (1988) argued that “understanding is not something that happens because of what a speaker says: the listener has a crucial part to play in the process, by activating various types of knowledge, and by applying what he knows to what he hears and trying to understand what the speaker means” (p. 6).

Listeners are not empty containers in which the information is
poured. They interpret the information they are receiving according to their individual characteristics, their background knowledge (knowledge about the topic and about the world), the context, and other non-linguistic variables (BUCK, 2001). Buck (2001, p. 8) explained that “when we listen, we use our background knowledge of the world to set up expectations, and then use those expectations to help us comprehend what we hear”. Those expectations might be different among listeners, as well as their motives for listening. Thus, listening is an individual process: different people have different interests and different needs. They “make different inferences, and they have different interpretations of the texts they hear” (BUCK, 2001, p. 29).

Understanding the characteristics of the spoken text which are different from written texts help us have a clearer idea of the listening construct. One characteristic of spoken text is that “speech is encoded in the form of sound” (BUCK, 2001, p. 4). There is no blank space or gaps between words in spoken language as there is in written texts. As Field (2019, p.11) explained, “the listener is presented with a string of syllables and has to work out where, within that string, one word ends and the next begins”. This process is called lexical segmentation. Also, the acoustic signal that listeners listen to may vary considerably due to, for instance, phonological modification, accent, stress and intonation.

Another characteristic of speech that differentiates it from written language is that “it is linear and takes place in real time, with no chance of review” (BUCK, 2001, p. 4). However, it is not always true that listeners have no chance of review. Nowadays, TV programs, for example, can be easily recorded or watched online, where listeners have the chance to listen again whenever they misunderstand something. Pilots are also encouraged to ask ATCOs to say again whenever they need clarification. Additionally, research has shown that learners can be taught to apply some strategic resources which can help them to become more effective listeners (O’MALLEY et al., 1989). Buck also argued that the speaker determines the speed of the text (which is usually fast, about three words per second), and that speakers often need to rely a lot on their memory. For this reason, he argued that processing needs to be automatic (a fast, effortless and unconscious process) rather than controlled (a conscious process which requires attention).

However, it is important to note that, nowadays, there are situations in which the listener determines the speed of the text (on YouTube, for instance, the speed of the video can be adjusted - it can be played at one quarter of the normal rate up to twice the normal rate). Another characteristic that differentiates spoken from written texts is that speech “is linguistically different from written language” (BUCK, 2001, p. 4). Spoken texts are usually unplanned; sentences are not clearly articulated. Speech usually consists of less complex idea units; vocabulary and grammar in spoken texts are usually simpler (although the use of idioms might make it more difficult for listeners) and less formal (grammatical structures might even be incorrect). Hesitations,
false starts and restatements are frequent. For all these reasons, Buck argued that testing listening is much more challenging than testing reading.

It is important to point out the difference between listening and hearing. While the verb *to hear* can be understood as “to receive or become conscious of a sound using your ears” (CAMBRIDGE DICTIONARY, n.d., Definition 1), the verb *to listen* is defined as “to give attention to someone or something in order to hear him, her, or it” (CAMBRIDGE DICTIONARY, n.d., Definition 2). A person could easily hear without listening, but never listen without hearing. Hearing is a precursor for listening and provides a basis for it (ROST, 2016). Although both hearing and listening involve physiological and neurological processes, listening is more intentional and involves a desire to understand what is being heard. This process involves both linguistic and non-linguistic knowledge (BUCK, 2001).

Linguistic knowledge includes phonology, lexis, syntax, semantics, and discourse structure. Buck explained that, although people tend to think that processing different kinds of knowledge occurs in a sequence that goes from the smallest sound segments (phonemes) to consecutive higher stages (lexis, syntax, and so on), it does not. “Listening comprehension”, as he states, “is a top-down process in the sense that the various types of knowledge involved in understanding language are not applied in any fixed order – they can be used in any order, or even simultaneously, and they are capable of interacting and influencing each other” (p. 3).

Non-linguistic knowledge refers to “knowledge about the topic, about the context, and general knowledge about the world and how it works” (BUCK, 2001, p. 2). When the listener knows the topic of the spoken text, they can use this knowledge to interpret what has been said; they can use their world knowledge in order to make inferences. Buck drew attention to the fact that the listening process is inferential and that background knowledge is important to listening comprehension:

If the listener shares the same knowledge as the speaker, much of what is being said can be understood by means of inferences based on shared background knowledge. However, if the listener has no knowledge relevant to the particular events being described in the text, then it will be more difficult to make inferences, and comprehension will be more dependent on interpreting the linguistic information. (BUCK, 2001, p. 20)

The context in which the communication takes place is also very important. There are different types of contexts. The context that is set by the text is referred to by psycholinguists as the *co-text* (BUCK, 2001). Sociolinguists consider the *context of the situation* in which the communication occurs to be very important. Another type of context that needs to be taken
into consideration is the **cognitive environment**. As listening comprehension happens in the mind of the listener, it is a cognitive process. According to Buck, “the context of interpretation is the cognitive environment of the listener” (p. 29), and it includes all the other contexts. For him, “it is the one that has the strongest influence on comprehension” (p. 21).

Furthermore, in order to better understand the listening construct, it is important to comprehend the difference between declarative knowledge and procedural knowledge (BUCK, 2001). Declarative knowledge is knowing something (e.g., knowing a grammatical rule), whereas procedural knowledge is knowing how to do something (e.g., knowing how to apply that grammatical rule). What seems to be important for listening is procedural knowledge, and that is what language testers should focus on (BUCK, 2001).

For listeners to understand a word, they have to access their mental lexicon to recognize the word and understand its meaning (BUCK, 2001). The acoustic information and the knowledge of the context are two kinds of information that are used in order to understand the word and its meaning. However, words are not usually understood in isolation, but as part of an idea unit, which may be understood as “short phrases or clauses ... strung together in a rather loose way, often connected more by the coherence of the ideas than by any formal grammatical relationship” (p. 9). Parsing is an important concept in this context. It involves establishing the relationship between the words and the utterances, with the help of syntactic and semantic clues. For this reason, as Buck (2001, p. 17) pointed out, “idea units are hardest to process when both the semantics and the syntax are challenging”. Nevertheless, listeners need not only to process idea units, but also, and very importantly, they need to process much larger linguistic units: they need to process connected discourse. At the discourse level, cohesion is an important variable that needs to be addressed by test developers. Understanding ideas connected by connectors, and determining what other cohesive devices (e.g., definite articles and pronouns) refer to is crucial (BUCK, 2001; FIELD, 2020).

The listening situation in which the listening takes place may also impact the listening process (BUCK, 2001). A communication between pilots and ATCOs, for example, is likely about traffic instructions, clearances, information about weather, airport, and so on. The listener’s role may be collaborative or non-collaborative. Non-collaborative listening requires only understanding what has been said, whereas collaborative listening requires “making appropriate requests for clarification, back-channelling, making responses to interactional language, or taking responsibility for organising turn-taking” (BUCK, 2001, p. 12). These skills are not usually considered to be listening skills. However, some listening situations surely make these kinds of demand on the listeners. Pilots most of the time need to be collaborative listeners, as they need, for example, to repeat the instructions they were given (read-back), and to ask for clarification whenever they do not understand what they heard. However, sometimes they also need to listen to recordings
(e.g., Automatic Terminal Information Service – ATIS). In cases like this, the recording keeps constantly playing, so they can hear it as many times they need.

Moreover, it is also important to consider the function of the interaction, which can be transactional, where the purpose is to communicate important information that needs to be understood by the listener (e.g., ATCO’s instructions), and interactional, where the purpose is just to interact socially (e.g., small talk). As Buck (2001) pointed out, “in most listening situations, there is both transactional and interactional language use, although one will usually be more dominant in any particular situation” (p. 14). In pilot/ATCO’s communication, any kind of unnecessary information is discouraged (e.g., greetings); transactional language is dominant. It is important to point out that the term interactional cannot be misinterpreted, as being able to interact effectively is very important in the context of pilot/ATCO communication. As a matter of fact, as explained earlier in this chapter, interactions is one of the six categories included in the ICAO rating scale.

An important theoretical model of the listening process was developed by Nagle and Sanders (1986). According to this adult listening model, the sensory or echoic memory captures the acoustic input and transfers it to the working memory, where it is processed by an executive processor, and then passed to the long-term memory, where it is synthesised with the implicit and the explicit linguistic knowledge, as well as with other knowledge. For Nagle and Sanders, “comprehension becomes more efficient as knowledge increases, processes become automatic, and experience confirms the reliability of the learner’s decoding, inferring, and predicting” (p.22). According to Buck (2001), although this process may be considered oversimplified, it does include some important elements. However, as Buck pointed out, it does not explain how meaning is represented in the memory. Meaning, as he explains, may be built up as propositions or mental models. The latter is the most common. Buck (2001, p. 29) argued that “mental models constitute a very important part of the cognitive environment and help determine how later parts of the text will be interpreted”.

Another important model of the listening process was presented by Field (2019). In his simplified cognitive model, the listening process consists of five different phases (or operations). The first phase is input decoding, when listeners “relate the sensations reaching their ears to the sounds of the target language” (p. 10). This decoding might be at the phoneme or syllable level, but it is important to emphasize that “larger units can influence the recognition of smaller ones” (p. 10). The next stage is the lexical search, when listeners put together the sounds they hear to form words. Then comes the parsing operation, when “the listener has to assemble a group of words into a syntactic pattern” (p. 11). As listening occurs in real time, we can say that it is a tentative process, as listeners need to constantly revise their understanding. These three initial phases are perceptual, whereas the last two (meaning
construction and discourse construction) are considered conceptual, as they do not involve language. Meaning construction involves interpreting the information which was obtained according to the listener’s knowledge and to the context. It is in the last operation, the discourse construction, that the listener puts everything together by connecting the new information to what they have previously heard.

A number of scholars have developed taxonomies to try to describe the sub-skills that are involved in the listening process (e.g., AITKEN, 1978; CARROLL, 1972; RICHARDS, 1983; VALETTE, 1977). Most of these taxonomies were mainly based on what Buck (2001, p.57) called “theoretical speculation”, whereas others were based on empirical research. Although these taxonomies may not be a thorough list of the sub-skills required in the listening process, they can be useful because they show some of its very important components and they may help test developers think of what they should include in their tests of listening comprehension.

In short, these taxonomies tell us that the listening construct is composed of a number of sub-components (listening is multidimensional) and that both the ability to extract the basic linguistic information and the ability to understand that information in a wider communicative context should be tested. Linguistic processing (phonology, stress, intonation, lexis, syntax, semantics and discourse structure) needs to be assessed, as well as the interpretation of the co-text, the context of situation and world knowledge (e.g., summarizing, making inferences, understanding sociolinguistic implications, understanding the speaker’s communicative purpose). However, these taxonomies “give no indication of the relative importance of individual skills, nor do they provide guidance on how they should be sampled for test construction” (BUCK, 2001, p. 59). In the next section, after having presented the main features of listening, I will talk about some issues regarding the assessment of this skill.

2.2. The assessment of listening

As Wagner (2014, p.1) pointed out, “the assessment of listening has historically been somewhat neglected and even overlooked in the language literature”. The three main approaches to language testing can help us better understand how the construct of listening has developed. They are: the discrete-point approach, the integrative testing approach, and the communicative testing approach (BUCK, 2001). The discrete-point approach is influenced by the behaviourist theory. According to this approach, the units of linguistic knowledge can be identified, isolated and tested separately. Lado (1961) was the main supporter of this approach. He believed that the assessment of listening should focus on the recognition of elements of oral language, including segmental phonemes, stress, intonation, grammar and vocabulary. The theoretical view of listening that is implicit in this approach, as Buck (2001, p.66) argued, is that comprehension is “understanding language on a local, literal level, and meaning is treated as something that is contained
within the text, and the listeners’ task is to get it out”.

The integrative testing approach, advocated by Oller (1979), highlights the importance of knowing how to use the language. This approach “puts the emphasis on assessing the processing of language as opposed to assessing knowledge about the elements of the language” (BUCK, 2001, p. 67). Somewhat similar to the integrative testing approach, the communicative testing approach also recognizes the importance of language use. However, the focus shifted from the importance of testing how to use the language to the importance of testing the use of language for the purpose of communication. As Buck (2001, p. 83) explained, the idea was that “testers should be less concerned with how much a person knows about the language, and more about whether they can use it to communicate effectively”.

An important characteristic of communicative testing is authenticity of texts and tasks (BUCK, 2001). Although some authors do not differentiate authentic texts from genuine texts, the difference between the two needs to be considered. Genuine texts are taken from the real world, and present some features of spoken language, such as assimilation, elision, hesitations, and reformulations. For some authors (e.g., FIELD, 2019), the use of genuine texts in listening tests is ideal.

However, using genuine texts also presents some disadvantages; for example, difficulty with copyright, sound quality problems, or a narrow range of texts types to choose from (ROSSI; BRUNFAUT, 2021). Genuine texts may be authentic if they reflect real life purposes. However, genuineness does not necessarily imply authenticity. Genuine texts are not authentic if they do not reflect real life purposes. Authentic texts might not be genuine; they can be created or adapted texts. What is important for test development is that the texts correspond to the target language use (TLU), no matter where they came from. TLU tasks can be defined as “a set of specific language use tasks that the test taker is likely to encounter outside of the test itself, and to which we want our inferences about language ability to generalize” (BACHMAN; PALMER, 1996, p. 44). For a test task to be authentic, the relationship test takers have with the text needs to reflect the TLU situation. As Rossi and Brunfaut (2021, p.17) pointed out, “while genuine texts certainly have their place in assessing listening, they might not always be suitable for item generation … since they might become inauthentic when taken out of their original context and might not allow for adequate language sampling to generalise test results to the TLU domain”. Authenticity has been the subject of much debate in the last two decades. As Ockey and Wagner (2018) pointed out, while some scholars fully embrace it, others are very reluctant to support it.

Although authentic tests may be more complicated to develop and administer, I believe that the use of more authentic tasks and texts can help to better assess the construct that we want to measure. As Ockey and Wagner (2018, p.3) explained, “by using test tasks that have many of the same characteristics as the target language use task, test users should be able to
make more valid inferences about the test takers’ ability beyond the testing context”. However, test developers need to keep in mind that authenticity does not automatically imply construct validity (FULCHER; DAVIDSON, 2007).

The communicative testing approach has had a profound influence on language testing, since its beginning until today. By understanding these three approaches, we can see “the development of an expanding view of the listening construct: from the narrow view of listening as recognizing elements, through listening as language processing, to the more current idea of listening as interpreting meaning in terms of a communicative context” (BUCK, 2001, p. 93). A brief explanation about the development of the assessment of listening by Cambridge ESOL may illustrate how the assessment of listening has developed: Cambridge ESOL, one of the major English test providers in the world, has been assessing listening since 1913, when its first English examination was launched (MILANOVIC; WEIR, 2013). Back then, listening was assessed through a dictation, part of the oral paper, as well as through a conversation, which was part of the same paper.

Additionally, there was a written paper on phonetics. It was a very demanding test. Since then, the way Cambridge test developers understand the construct of listening has developed significantly. In the 1970s, listening started to be tested in a specific listening comprehension test. Listening was only assessed in isolation, not as part of an interaction. At that time, test-takers had to listen to literary passages being read in the examination room “requiring not much more than comprehension of factual detail and utterance-level processing” (MILANOVIC; WEIR, 2013, p. xi). This was, of course, problematic for many reasons (e.g., issues with test standardization). With the advent of communicative approaches to language teaching and testing in the 1970s, tasks started to reflect more the real-world tasks.

As Field (2019, p.1) argued, “testing second language listening proficiency validly and reliably has always posed a challenge”. The challenges faced by test developers when writing test specifications for listening tests include deciding whether the candidates will be allowed to listen to the text more than once (TAYLOR; GERANPAYEH, 2011), issues related to task authenticity (BRINDLEY, 1998; LYNCH; MENDELSON, 2010; WAGNER, 2014) and to memory (WU, 1998). It is also important to consider the complexities of cognitive processes involved in listening so that “the cognitive processing activated in the test taker by a test task corresponds as closely as possible to what they would expect to do in the (...) listening context” (TAYLOR; GERANPAYEH, 2011, p. 96).

Listening test developers must try to understand very well what language they are targeting so they can develop a test which is similar and representative of that domain (BACHMAN; PALMER, 1996; WAGNER, 2014). Identifying the characteristics of the specific situation in which the language is being used is of utmost importance. They need to think carefully about the purpose of the test and the specific target language situation in order to make
informed decisions. For example, if the TLU domain involves listening and speaking together, assessing listening at the same time as speaking may be necessary. The recognition of the importance of integrative testing questions the aforementioned traditional four skills approach, which assesses reading, writing, speaking and listening independently (OCKEY; WAGNER, 2018).

Assessing listening together with speaking may be challenging, “but it is a necessary and advisable goal” (WAGNER, 2014, p. 6). Field (2020) also highlighted the importance of assessing interactive listening. He argued that the listening processes that are employed by listeners in an interactive conversation might be different from the processes that occur when listening in isolation, and that interactive listening is more cognitively demanding than listening to a recording. Moreover, as Lam (2021, p.20) argued, interactive listening is “fundamentally social” and “needs to be assessed outside the boundaries of (receptive) listening tests”.

In the context of pilot/ATCO communication, most of the listening performed by pilots is in an interaction with the ATCO, not listening to recordings (e.g., ATIS). Therefore, I believe that in this context, for both pilots and ATCOs, it is extremely important to assess integrative listening. As Green (2017, p.8) argued, “air traffic controllers (ATC) need to be able to demonstrate not only good listening skills but also the ability to interact when communicating with pilots or fellow ATC colleagues … Therefore, an interactional listening task is likely to have much more validity”. The ICAO LPR test design criteria (ICAEA, 2021) highlight the importance of assessing listening in isolation. They say that “it is possible for tests to also evaluate comprehension subjectively in an interactive context in addition to having a dedicated listening test section, but not to the exclusion of including dedicated listening comprehension test sections” (ICAEA, 2021). However, I believe they emphasize that listening should be tested separately from speaking, in a test that is entirely dedicated to the assessment of listening, without drawing much attention to the importance of the assessment of integrative listening. In my opinion, this should have been highlighted, not just as a possibility, but as a necessity.

3. FURTHER IMPLICATIONS FOR THE ASSESSMENT OF PILOTS’ LISTENING IN THE CONTEXT OF RADIOTELEPHONY COMMUNICATIONS

As Green (2017, p. 29) pointed out, “the definition of what listening is will differ according to the purpose of the test and also the target test population”. Radiotelephony communications between pilots and ATCOs are very different from normal daily conversations. This kind of communication “represents a very specialized and socially significant form of discourse” (READ; KNOCH, 2009, p. 21.3). These communications consist mostly of standard phraseology, which is a sub-language for routine situations, such as orders, requests, advice, permissions, approvals, etc. It was created as an attempt to standardize the language use in pilot/ATCO radiotelephony communications in an effort to avoid misunderstandings. Whenever there is
no standard phraseology to convey meaning, pilots and ATCs rely on plain English, which was defined by ICAO as “the spontaneous, creative and non-coded use of a given natural language” (ICAO, 2010, p. x).

An important characteristic that differentiates listening to pilot/ATCO communications from other types of listening is that the topics are related to aircraft operation. Although these topics are limited to situations that are relevant to the operation of the aircraft, its range can vary considerably. For a list of communicative language functions, events, domains, subdomains and tasks associated with pilots/ATCOs communications, test developers may refer to Appendix B of ICAO’s DOC 9835 “Manual on the implementation of ICAO Language Proficiency Requirements” (ICAO, 2010), which was based on research at the Direction Générale de l’Aviation Civile (France) as well as on Ramos et al. (1999). However, it is important to point out that this list has been considered vague and insufficient to be used in test development (ARAGÃO; SCARAMUCCI, 2020). Aragão and Scaramucci suggest that this list needs to be informed and validated by subject matter experts so that it can truly reflect the pilot/ATCO work environment.

Some features of pilot/ATCO communications may help their communication to be more effective, while others may make it more challenging. The features that may contribute to successful communications include: absence of overlapping (as pilots and ATCOs talk via radio), the fact that messages should be grammatically simple, meaningful and brief (PRINZO; BRITTON, 1993), and the possibility to ask the speaker to repeat or rephrase the message whenever necessary. In pilot/ATCO communications, it is advisable to ask for clarification, differently from the listening situation that Buck (2001) and Lynch and Mendelsohn (2010) described, where listeners usually have just one opportunity to listen to an input. In this context, asking for repetition on clarification is a matter of safety.

Although “these extra communications increase radio frequency congestion and reduce the efficiency of information transfer” (PRINZO; BRITTON, 1993, p.1), checking information, seeking confirmation, and requesting clarification need to be encouraged in both teaching and testing. Allowing a test taker pilot to ask for clarification, or to listen to the input again, is, in my opinion, a good practice, as it corresponds to their TLU domain.

Some features of aeronautical English that may contribute to comprehension problems are:

**Absence of visual support** – Communications between pilots and ATCOs is a voice-only interaction, with no visual references. In other types of interactions, messages are usually conveyed not only by the sound, but also by the use of non-verbal signals (BUCK, 2001; TANENHAUS et al., 1995; WAGNER, 2013). As visual support may help listeners understand spoken texts, its lack may be considered a difficulty factor in the interaction between pilot and ATCOs. As they are not able to see each other, test tasks to assess pilots’ listening should not include audiovisual texts of ATCOs. However, pilots
also talk to ATCO when they are at airports, looking at the taxiways, runways, airport signs, other aircraft, etc. Therefore, test developers may include images that help to contextualize the task (e.g., a controller may ask a pilot to check if he can see oil leaking from another aircraft that is close by, so the task may include a picture of the aircraft, as seen from the test taker’s cockpit perspective).

**Different accents** – Accent plays an important role in understanding spoken texts. As pointed out by Buck (2001), “when listeners hear an unfamiliar accent (…) this can cause problems and may disrupt the whole comprehension process. An unfamiliar accent can make comprehension almost impossible for the listener” (p.35). Research has shown that even native speakers process information slower when listening to unfamiliar native accents under adverse listening conditions (ADANK et al., 2009). Pilots seem to spend more time listening to ATCOs who are non-native speakers of English than to native speakers (GARCIA; FOX, 2020; ICAO, 2010). They are exposed to a wide variety of familiar and unfamiliar accents. Although Level 4 pilots and ATCOs are supposed to assume a dialect and/or accent which is intelligible to the aeronautical community, their accent might sometimes interfere with ease of understanding (see descriptors for pronunciation in the ICAO rating scale). As listening to unfamiliar accents is an element that seem to contribute to listening comprehension difficulties in pilot/ATCO communications (GARCIA; FOX, 2020), the ability to understand multiple speech varieties should be part of the listening construct of pilots/ATCOs’ aeronautical English tests. Harding (2018) also argued that in some assessment constructs, such as in the air traffic control (ATC) domain, “the ability to cope with an unfamiliar accent could be conceptualized as a central part of the listening construct” (p. 97).

**Emotional stress or increased mental workload** - As there is standard phraseology set for all routine situations in pilot/ATC radiotelephony communications, the need to use plain English arises usually in non-routine, abnormal or unexpected situations (e.g., very bad weather, heavy workload, time pressure and mechanical failure). When dealing with these kinds of situations, pilots and ATCOs may get emotionally stressed (ALDERSON, 2009), and this stress may compromise their linguistic performance. Research has shown that increased workload may affect pilot’s accuracy when reading back ATCO’s transmissions (ESTIVAL; MOLESWORTH, 2016).

**Problems related to the environment and the channel** – Other elements that may influence listening comprehension difficulty in pilot/ATC radiotelephony communications are related to the environment and the channel. Cockpits are noisy environments, especially because of the loud noise produced by the engines (MOLESWORTH, 2016). In addition, radio transmission is frequently poor. Research has shown that pilots think it is challenging to interact with the ATCOs via radio, as many times they need to use clarification strategies, not because of linguistic problems, but because of the bad quality of the transmission (ESTIVAL; MOLESWORTH,
However, this seems to be less of a problem nowadays because of the improvement in the quality of the transmissions (GARCIA; FOX, 2020).

**Speech rate** - Speech rate is also a factor that might increase difficulty for pilots and ATCs to understand each other. Buck (2001) pointed out that research indicates that “the faster the speech, the more difficult it is to comprehend” (p. 38). Although ICAO recommends ATCOs to speak at a rate of 100 words per minute, research has shown that they speak at a much faster rate (GARCIA; FOX, 2020; PRINZO; BRITTON, 1993).

**Threats of intercultural factors** – The intercultural context in which communications between pilots and ATCOs take place may contribute to misunderstandings. Factors related to power distance, communication styles and non-collaborative behaviour, among others, may negatively affect the effectiveness of communications (MONTEIRO, 2019).

**Unfamiliar topics** - Aviation professionals may come across a wide variety of unexpected or unfamiliar topics and this might have a negative impact on their ability to understand what they hear. Standardized phraseology consists of a reduced vocabulary of around 400 words (ICAO, 2010). However, as previously discussed, pilots and ATCs need to use plain English to deal with abnormal and emergency situations. In this case, the range of vocabulary that may be used is extensive. Research has shown that topic familiarity impacts language listening comprehension, regardless of the listener’s proficiency level (SCHMIDT-RINEHART, 1994). It is also worth noting that, according to Révész and Brunfaut (2013), lexical complexity has a significant impact on listening difficulty. They argued that a low proportion of function words, a high frequency of academic words, greater lexical density and wider lexical diversity contribute to the need of an increased processing effort to understand speech.

Understanding these features of radiotelephony communications may help test developers define the listening test construct. These features may be included in the test tasks so that they reflect the TLU. For example, test developers need to consider the speech rate of the spoken texts, the range of accents to be included, and also the possibility to add aircraft background noise to the recordings performed by pilots and ATC room background noise to recordings of ATCOs (some interference may be included too). As Douglas (2000) argued, tasks of language for specific purposes (LSP) tests need to be based on the characteristics of the TLU domain. According to him, “it is this analysis of target language use task characteristics which will allow us to make inferences about language ability in the specific purpose domain” (p. 14).

Miscommunication in ATC might occur due to problems related to the speaker, problems related to the channel, and/or problems related to the listener (ICAO, 2010). The problems related to the speaker include propositional failure (e.g., inaccurate assumptions about shared background knowledge with the listener), encoding failure (e.g., wrong choice of
vocabulary or grammar mistakes), or delivery failure (e.g., pronunciation problems or inappropriate speech rate). The problems related to the listener can be due to decoding failure (e.g., language, attention, memory problems), to interpretation failure, which can be a consequence of the speaker encoding problems, and to feedback failure.

Cushing (1994) classified pilot/ATCO communication problems into language-based problems and problems not based on language. For him, the communication problems based on language which he identified were: problems of language (e.g., ambiguity, homophony, intonation), problems of reference (e.g., uncertain reference, uncertain addressee), problems of inference (e.g., implicit inference, lexical inference, unfamiliar terminology, false assumptions), and problems involving repetition (e.g., partial readbacks). The communication problems not based on language which he lists are: problems with numbers, problems with radio, problems of compliance, and other general problems.

As forementioned, test item writers also need to take into consideration issues related to memory. Human memory plays an important role in comprehending what we hear (ORTEGA, 2009; WU, 1998). As previously discussed in this chapter, human working memory capacity is limited and “when the task demands are high, as in a test of listening comprehension, often because of both storage and processing needs, the computation will slow down and some partial results from working memory processing may be forgotten” (WU, 1998, p. 23). According to Clark et al. (1990), for example, when pilots’ working memory needs to process transactions with a higher number of speech acts (phrases or clauses), there is a higher probability of misunderstanding. As Buck (2001) pointed out, the limit to the capacity of the working memory “seems to be restricted to about seven units, or chunks of information” (p. 77). Garcia and Fox (2020)’s study suggested that the average number of pieces of information in ATCO’s messages is three or four, but as real-world transmissions may be longer than that, some test recordings might also include more items. If the test task demands too much of the test takers’ memory, the inferences we can make about the test takers’ listening ability might not reflect their real ability. Therefore, it is important that aeronautical English test developers take into consideration the limitations of human memory in order to ensure greater construct validity.

Another important issue that test developers need to consider is whether test takers should be allowed to take notes while listening to the spoken text. Pilots are used to taking notes of ATCO’s instructions, so it is natural for them to want to take notes. As discussed by Hughes (2003, p. 167), “where the ability to take notes while listening to, say, a lecture is in question, this activity can be quite realistically replicated in the testing situation”. From my own experience applying the Santos Dumont English Assessment (SDEA), the Brazilian national examination for testing pilots’ English proficiency, I can say that pilots get stressed and anxious if they are not allowed to take any
notes during the test. Pilots were not allowed to take notes during SDEA from 2007 to 2010. They used to complain that the test was assessing their memory capacity rather than their listening skills. Since 2010, they have been allowed to take notes while listening to the audio texts and this change in the testing procedures was very appreciated by the test takers.

It is important to note that, although plain English should only be used by pilots and ATCOs when the standard phraseology does not suffice, it is frequently used when phraseology would be enough (KIM; ELDER, 2009; HOWARD, 2008; MORROW et al., 1994). As a matter of fact, research has indicated that 70% of the communication between pilots and ATCOs does not comply with the prescribed phraseology (MELL, 1992, as cited in ICAO, 2010). According to ICAO (2010), “users, particularly expert speakers of a language, for all sorts of respectable reasons such as pressure of work, and less respectable reasons such as carelessness and insensitivity, fail to adhere to prescribed ICAO standardized phraseology, thereby creating possibilities for misunderstanding in a busy international environment” (p. 3-5).

Hamzah and Fook Fei’s study (2018) found that one of the main factors that causes miscommunication between pilots and ATCOs is incorrect utilization of phraseology. Test developers need to face the challenge whether to include in their listening tests only texts in which phraseology is used appropriately and in accordance with the ICAO standardized phraseology, as recommended by ICAO (2010), or to include texts with certain deviations from phraseology (e.g., “thirty minutes” instead of “three zero minutes” or “point” instead of “decimal” for frequencies). This online forum comment can illustrate this issue:

Get rid of the word “decimal” in ATC frequency changes. ATC will use phrases like contact ground “point” 75 or contact tower on 118 “point” 3. Pilots don’t even use the “point” for most freq changes and just say something like 24 4 for 124.4. Tower controllers will usually just say contact departure expecting you to know or they will say contact departure on 124 “point” 7. The “Point” being is that if you listen on any freq on LiveATC you will not hear many if at all any controllers or pilots use the word “decimal”. It takes up too much bandwidth. And trying to remember the freq numbers when they throw in that 3-syllable word between them is much more difficult. So Asobo wanted to make the scenery real why not make some of the ATC a little more real? In all of my flying years I have rarely if ever heard or used that word. (WHATACROCKTHIS, 2020)

I particularly believe that if the purpose of the test is to assess the ability to understand real-life communications, unstandardized phraseology may be included in the test, but care needs to be taken. The test rubrics must be clear
in relation to what is going to be tested and why, as well as what to expect from the test takers. This means that if the recordings include communications that do not comply with ICAO standardized phraseology, test takers should be made aware that they might encounter non-standardized communications in the test because unfortunately this is what they should expect to hear in real-life ATCO communications. They should also be encouraged to use standardized phraseology when responding to the test tasks. It is important to emphasize that pilots and ATCOs need to be trained to use standardized phraseology as much as possible in order to enhance safety.

4. CONCLUSION

As Field (2019) argued, having extensive information about the nature of the skill we want to assess “enables us to shape the material and tasks that we devise so as to ensure that they truly measure competence in the skill” (FIELD, 2019, p. 7). There is still a lot to be known about the nature of listening and about its operationalization in language tests, especially in the context of pilot/ATCO communications. Nevertheless, it is clear that for the assessment of languages for professional purposes, “the theoretical construct and its operationalization should, at least in part, be derived from or be relevant to the TLU domain” (KNOCH; MACQUEEN, 2020, p. 190).

To conclude, listening is complex, especially in a second or foreign language. The assessment of listening is a challenging endeavour. Test developers need to make the best they can to develop tests and create tasks that truly measure the construct that they seek to assess, especially in high stakes testing contexts such as the pilot/ATCO’s communications.

ACKNOWLEDGEMENTS

I would like to express my gratitude to DECEA and ICEA for publishing this book, especially GEIA’s leader, Patricia Tosqui Lucks, for her hard work organizing GEIA. I would also like to thank all GEIA members, my thesis supervisors and advisory committee, as well as my colleagues, friends and family for their continuous support. Most importantly, I thank God, because without Him writing this chapter would not have been possible.

REFERENCES


FIELD, J. Rethinking the Second Language Listening Test – From Theory to Practice. Equinox, 2019.


FOX, J. Introduction. In J. Fox et al. (Eds.), Language testing reconsidered.

GARCIA, A. C. M. 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. 2015. Master's Dissertation in Language Testing - Department of Linguistics and English Language, Lancaster University, UK, 2015.

GARCIA, A. C. M.; FOX, J. Contexts and constructs: Implications for the testing of listening in pilots’ communication with air traffic controllers. The ESPlenialist, 41(4), 1-33, 2020. https://doi.org/10.23925/2318-7115.2020v41i4a4


KIM, H.; BILLINGTON, R. Pronunciation and comprehension in English as a lingua franca communication: Effect of L1 influence in international aviation communication.


WHATACROCKTHIS. Get rid of the word “decimal” in ATC frequency changes. ATC will use phrases like contact ground “point” 75 [Comment on the online forum post ATC using “decimal” when giving freq. change]. August, 2020. Microsoft Flight Simulator. https://forums.flightsimulator.com/t/atc-using-decimal-when-giving-freq-change/212705

Chapter 4: Manuscript 3 – Investigating the construct of aeronautical English listening testing: A qualitative analysis of the ICAO rating scale

The manuscript in this chapter comprises a journal article published in *The Journal of Teaching English for Specific and Academic Purposes* in April 2023. The journal’s Editor-in-Chief is Dr. Nadežda Stojković, University of Niš, in Serbia. The guest editors were Vanya Katsarska, from Bulgaria’s Air Force Academy, and Dr. Anna Borowska, from the University of Warsaw, in Poland. The purpose of the study presented in this article was to analyze the descriptors of comprehension in the ICAO rating scale in order to better understand the listening construct underpinning the ICAO policy and to examine the descriptors for the other criteria in the rating scale in order to find out how they could help inform the definition of the listening construct of an aeronautical English listening test.

Alternatively, the article can be read through the following hyperlink:


[https://doi.org/10.22190/JTESAP230220007G](https://doi.org/10.22190/JTESAP230220007G)

The following corrections/suggestions of improvement should be acknowledged:

<table>
<thead>
<tr>
<th>Article page number/line</th>
<th>From</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>72/35</td>
<td>(see Alderson 2011; Aragão 2018)</td>
<td>(see Aragão 2018; Alderson 2011)</td>
</tr>
<tr>
<td>75/7</td>
<td>in the Annex 1</td>
<td>in Annex 1</td>
</tr>
<tr>
<td>75/17</td>
<td>However, when discussing the results</td>
<td>When discussing the results</td>
</tr>
<tr>
<td>78/28</td>
<td>questions5, 6</td>
<td>questions 5, 6</td>
</tr>
<tr>
<td>-------</td>
<td>---------------</td>
<td>----------------</td>
</tr>
<tr>
<td>79/35</td>
<td>This frequency may</td>
<td>Pronunciation may</td>
</tr>
<tr>
<td>81/24</td>
<td>Interference of pronunciation, stress, rhythm, or intonation, a linguistic complication on ease of understanding</td>
<td>Interference of pronunciation, stress, rhythm, or intonation, on ease of understanding</td>
</tr>
</tbody>
</table>
INVESTIGATING THE CONSTRUCT OF AERONAUTICAL ENGLISH LISTENING TESTING: A QUALITATIVE ANALYSIS OF THE ICAO RATING SCALE

Angela C. de M. Garcia
Carleton University, Canada

Abstract. Since the publication of the International Civil Aviation Organization (ICAO)’s language proficiency requirements, a number of different tests have been developed and implemented around the world to assess pilots and air traffic controllers’ proficiency in English. Meanwhile, researchers have questioned the clarity and appropriateness of the policy, and the reliability of tests (e.g., Alderson 2011; Douglas 2004; Emery 2014). ICAO has recently acknowledged that, over the years, multiple interpretations of the policy have led to practices that might undermine the meaningfulness of aviation English tests (ICAO 2022b). However, the ICAO Rating Scale remains as the instrument to be used in assessing pilots and air traffic controllers’ aeronautical language proficiency. Thus, this article explores the construct of aeronautical English listening tests stated in the comprehension descriptors of the ICAO rating scale, as well as the elements of the other descriptors that may inform the definition of this construct. An in-depth content analysis of the rating scale was conducted by using the “interview technique”, as described by O’Leary (2021). Results provide useful information for the development of listening tests in the aeronautical context. A better interpretation of the construct informed by the policy can help to reduce the differences among test implementations around the world and further contribute to more standardized and meaningful testing practices.

Key words: language testing, ICAO rating scale, listening comprehension, construct definition

1. INTRODUCTION

In 1998, after a tragic accident that led to 349 fatalities in 1996, India submitted a working paper asking the International Civil Aviation Organization (ICAO) Assembly to consider the lack of language proficiency of pilots and air traffic controllers (ATCOs) with a high degree of priority (Popa 2019). In 2003, the ICAO Council adopted the Amendment 164 to the Annex 1 (entitled Personnel Licensing) to the Convention on International Civil Aviation, requiring pilots and ATCOs to “demonstrate the ability to
speak and understand the language used for radiotelephony communications” (ICAO 2022a: paragraph 1.2.9.1). For pilots and ATCOs to demonstrate their speaking and listening abilities, they need to be tested, unless they are native speakers of English. Appendix 1 to Annex 1 includes a set of holistic descriptors, which describe five abilities which pilots and ATCOs should demonstrate, and, Attachment A includes the ICAO Language Proficiency Rating Scale, which details six analytic criteria (pronunciation, structure, vocabulary, fluency, comprehension, and interactions) for the six proficiency levels mentioned in paragraph 1.2.9.1 (level 1, Pre-elementary, to level 6, Expert). For pilots and ATCOs to meet the ICAO LPRs, they must demonstrate compliance with the holistic descriptors and the ICAO Operational Level (Level 4), detailed in the rating scale. In other words, to be allowed to fly internationally or to control international flights, pilots and ATCOs should be awarded at least level 4 in all six criteria of the rating scale. To help civil aviation authorities and testing service providers to develop tests to assess pilots and ATCO’s aeronautical language proficiency, ICAO published, in 2004, the first edition of the DOC 9835 – Manual on the Implementation of ICAO Language Proficiency Requirements (ICAO 2004). In that time, Douglas argued that in the aviation English context it would be very important to have a clear picture of the language which 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” (Douglas 2004: 250). After the publication of the ICAO language proficiency requirements (LPRs), the so-called aviation English tests (which I call aeronautical English tests [see Tosqui-Lucks and Silva 2020]) started to be developed around the globe by different organizations. In 2010, Alderson (2010) conducted a survey on aviation English tests for pilots and ATCOs and concluded that “little or no confidence can be had in the meaningfulness, reliability and validity of several of the aviation language tests currently available” (p. 63). He then interrogated if the ICAO scales are explicit, relevant, and adequate.

In 2016, during the 39th session of the ICAO Assembly, Brazil (2016) presented a working paper inviting the Assembly to review the ICAO LPRs, arguing that the construct underpinning the policy is unclear and under-represented. The response to this working paper was that “it [the working paper] did not present sufficient evidence that existing language proficiency requirements posed a safety threat”, and that “the need to revise the language proficiency requirements could be considered once additional implementation data was [sic] collected through the different initiatives of ICAO” (ICAO 2016, para. 35.155). Unfortunately, over the years, as ICAO itself remarked, multiple

---

1 A note to the ICAO Annex 1 paragraph 1.2.9 explains that the language to which paragraph 1.2.9 refers may be English or the language “normally used by the station on the ground” (ICAO 2022a, p. 1-17). In practice, the language that is commonly evaluated is English, because when pilots and ATCOs use the language used by the station on the ground, they usually use their native language. As another note to paragraph 1.2.9 informs, pilots and ATCOs who demonstrate expert language proficiency (e.g., native speakers) do not need to be formally evaluated.

2 Due to a limitation on the number of words of this article, the holistic descriptors and the rating scale were not included in the appendix, but Annex 1 is available at https://elibrary.icao.int/home.

3 A construct can be understood as “the theoretical entity that the test developers and test users intend the test to measure, the quality or qualities of the test takers we wish to make inferences about” (Douglas 2010: 33).
interpretations of the policy have led to practices that might have undermined the meaningfulness of aviation English tests (ICAO 2022b). The lack of clarity, appropriateness, and fairness of the ICAO policy, as mentioned by many (e.g., Douglas 2004, 2014; Emery 2014; Kim and Billington 2016; Kim and Elder 2015; Knoch 2014; Read and Knoch 2009), most likely contributed to such consequences. In spite of the criticism towards the policy, the LPRs have not changed and are still what Contracting States need to comply with, unless ICAO is notified of any differences between national regulations and the ICAO requirements. Thus, as Emery (2007) points out, “every testing programme and test instrument developed to measure the language proficiency of aviation operations personnel will employ the Rating Scale and Holistic Descriptors in each of the 190 ICAO member states” (p. 1). Therefore, test developers need to have a clear and deep understanding of the ICAO policy. Knoch and Macqueen (2020) argue that policy analysis “is a crucial aspect of work that needs to be completed before an assessment is developed or adopted” (Knoch and Macqueen 2020: 87).

Many aspects related to the policy could be explored. As Wodak (2006) points out, “there are obviously many relevant research issues and a variety of genres and public spaces where a precise linguistic analysis of oral, visual, or written texts will provide differentiated knowledge on aspects of language politics/policies” (Wodak 2006: 170). The focus of the present study is to investigate one specific aspect of the ICAO policy: the construct of the listening in isolation test represented in the ICAO rating scale. Criterion 3 of the ICAO test design guidelines4, which were developed by the International Civil Aviation English Association (ICAEA) in partnership with ICAO, recommends that “test instruments need to contain tasks dedicated to assessing listening comprehension separate from tasks designed to assess speaking performance” (ICAEA n.d.). It is necessary to assess listening in isolation in order to minimize a major threat to the validity of the interpretations and uses of tests scores: construct irrelevant variance, “that is, the test is too broad and contains excess reliable variance associated with other distinct constructs as well as method variance making items or tasks easier or harder for some respondents in a manner irrelevant to the interpreted construct” (Messick 1989: 14). ICAEA points out that “assessing comprehension at the same time as speaking compromises the validity of the result for comprehension”, and that “test developers need to be mindful of ensuring interference of ability in other skills do not unfairly influence the assessment results” (ICAEA n.d.). The ICAO test design guidelines adds that “this means test-takers are required to listen to prescribed recordings and then complete follow up comprehension tasks. Such tasks could be on paper, require test-takers to summarise information or answer prescribed written questions asked orally or provided on a test paper/computer screen” (ICAEA n.d.). It is important to point out that, although it is necessary to assess listening in isolation, the assessment of interactive listening is also essential in this context, as most of the listening performed by pilots and ATCOs happen as part of an interactive conversation between them. Field (2020) argues that the listening processes employed in conversations might be different from those employed when listening to a recording, and also more cognitively demanding. Lam (2021) also

---

4 At the time this article was written, this author was participating in the meetings of the ICAO Exploratory Group - Language Proficiency Requirements (EG-LPRs/03), which was created to revise the ICAO test design guidelines in order to have them published by ICAO as a handbook.
emphasizes that interactive listening “needs to be assessed outside the boundaries of (receptive) listening tests” (Lam 2021: 20). However, the focus of the present study is on the construct of a test that aims to assess listening in isolation. Hence, the research questions that this study addresses are the following:

**Research Question 1 (RQ1):** What is the listening in isolation construct represented in the descriptors for comprehension detailed in the ICAO Language Proficiency Rating Scale?

**Research Question 2 (RQ2):** How may the descriptors for the other criteria help to define the construct of a test to assess listening in isolation?

2. LITERATURE REVIEW

The ICAO rating scale was developed by a committee appointed by ICAO known as the Proficiency Requirements in Common English Study Group (PRICESG). ICAO (2010) describes that “this study group brought together, from Contracting States and international organizations, operational and linguistic experts with backgrounds in aviation (pilots, air traffic controllers and civil aviation authority representatives), aviation English training and applied linguistics” (para. 1.4.2). Estival, Farris, and Molesworth (2016) observe that this study group did not include language testing expertise. McNamara, Knoch, and Jason (2019) adds that this group “was dominated by representatives from English-speaking nations” (McNamara, Knoch, and Jason 2019: 19).

The ICAO policy has been criticized from a number of different perspectives. One of the main criticisms is related to the fact that native speakers do not need to be formally evaluated. Many authors (e.g., Borowska 2017; Douglas 2014; Estival, Farris, and Molesworth 2016; Kim 2013; McNamara, Knoch, and Jason 2019; Monteiro 2019; Read and Knoch 2009; Trippe 2018) argue that native speakers should have their ability to communicate evaluated. This test should assess their ability to accommodate to non-native speakers when necessary by, for instance, using simpler vocabulary or by speaking at a slower rate. I believe the main reason why Brazil’s paper to ICAO was not accepted was because the working paper highlighted that it was necessary to assess native speakers of English. Having the paper accepted would mean that some countries would have to spend money and time to assess and train their pilots and ATCOs. Asking pilots to pay a two-dollar fee to receive a new license with an “English Proficient” endorsement, as the United States were doing (Alderson 2011) would not be enough, and this is probably why the United States voted against the Brazilian request. As it has been argued, the politics involved in this context have a huge impact on the decisions that are made, and their agenda is often hidden (see Aragão 2018; Alderson 2011).

Another recurrent criticism to the ICAO policy lies in the fact that the guidelines advice that the focus of the language assessment should be on plain English proficiency, without taking into consideration the incorrect use of phraseology or the lack of technical knowledge of operations (ICAO 2010). DOC 9835 does mention that it is important to adhere
Investigating the Construct of Aeronautical English Listening Testing

73

to ICAO standardized phraseology. It also acknowledges that 70% of the radiotelephony speech acts do not comply with it. However, although a somewhat contradicting note to Annex 1’s Appendix 1 says that “the language proficiency requirements are applicable to the use of both phraseology and plain language” (ICAO 2022a), the policy does not require phraseology to be assessed. DOC 9835 explains that “it is acceptable that a test contains a scripted task in which phraseology is included in a prompt, but the test should not be designed to assess phraseology” (ICAO 2010: 6.3.2.9). Korean pilots and ATCOs who participated in Kim’s (2013) study believe that the non-observance of radiotelephony conventions plays a more important role on safety than proficiency in plain English. Indeed, underutilization of phraseology has shown to increase problems in communication (e.g., Howard 2008). Additionally, DOC 9835 says that “the test should not be designed to evaluate technical knowledge of operations” (ICAO 2010: 6.3.2.10). This is understandable, as the tests should not ask questions such as “What are the separation minima for aircraft being vectored for an ILS approach?” or “Describe the different flight modes of the A320 flight control system?” (ICAO 2010: 6.3.2.10). However, from analysing indigenous assessment criteria (the criteria adopted by domain language users to assess the effectiveness of communication [Jacoby and McNamara 1999]), Aragão (2018) argues that ATCOs consider non-linguistic elements, such as psychological aspects and operational knowledge, to be significant contributors to effective communication. McNamara, Knoch, and Jason (2021) also point out that “experienced pilots and air traffic controllers know that technical knowledge is an inextricable part of language use” (McNamara, Knoch, and Jason 2021: 17). Research findings (e.g., Kim 2013; Knoch 2009, 2014; Aragão 2018; McNamara, Knoch, and Jason 2021) support Douglas’s (2001) notion of specific purpose language ability in Language for Specific Purposes (LSP) testing, which he defined as “a construct that results from the interaction between specific purpose background knowledge and language knowledge” (Douglas 2001: 50). Unfortunately, it seems that, as Knoch, Deygers, and Kharnboonruang (2021) point out, when the ICAO rating scale was developed, indigenous assessment criteria were not taken much into consideration.

Not much is known about the development and validation of the ICAO rating scale (Kim and Elder 2009; Knoch 2009) and few studies have been conducted to investigate the construct represented in the ICAO rating scale. Knoch (2009) conducted a validation study to investigate how test developers, administrators, and raters viewed the rating scale. Although participants responded that they were generally satisfied with the rating scale descriptors, their responses to open-ended questions indicated a range of problems that stakeholders identified in the ICAO scale. For comprehension, the most cited problem was that “comprehension could not be accurately measured in a scale designed to assess speaking performance” (Knoch 2009: 31). This confirms Pfeiffer’s (2009) findings. Pfeiffer (2009) investigated inter-rater reliability in a German speaking test, and found it to be low, the lowest being for comprehension. She points out that “the ICAO descriptors are often incomplete and therefore need amendment, however with comprehension the user could easily have the impression 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” (Pfeiffer 2009: 56). She adds that “the wording of the level descriptors for comprehension is not very enlightening. According to my judgement, they are possibly the least well thought out in the entire rating scale” (Pfeiffer 2009: 57).

Similarly to Knoch (2009), Garcia (2015) interviewed very experienced ICAO LPRs test developers and raters to investigate their perceptions on the ICAO policy in general and, more specifically, on the ICAO rating scale. She reported a number of recurring themes criticized
by participants, such as the presence of contradictions within the policy, the lack of fit between the policy and the target language use (TLU) domain, the need to assess pilots and ATCOs ability to communicate effectively (not only proficiency in plain English), and the importance of following standard phraseology. Participants were asked to discuss the strengths and weaknesses of the ICAO rating scale descriptors for each criterion. One participant complained about the fact that comprehension is only one category out of six in the rating scale. He argued that “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”. Indeed, according to Feyten (1991), as cited by Buck (2001), it seems that people spend about 45% of the total time of communication using listening skills. The ICAO requirement itself emphasizes this importance by saying, as mentioned, that pilots and ATCOs must demonstrate their ability to do two things: speak and understand the language used for radiotelephony communications. The same participant in Garcia (2015)’s study advocated that comprehension “is an extremely important, if not more important, part of the overall proficiency construct in this case (Garcia 2015: 38). For him, comprehension should have its own rating scale. This belief is upheld by Knoch (2009) when she points out, as mentioned, that using a scale that was designed to assess speaking in order to assess comprehension is problematic. However, differently from Knoch’s (2009) results, that suggest that comprehension of cultural subtleties seems irrelevant, participants in García’s (2015) study pointed out that the main strength of the comprehension descriptors was this reference to comprehension of cultural subtleties. One of the participants even argued that this reference should have been included in other levels of proficiency, not only in level 6. Monteiro’s (2019) findings uphold García’s (2015), in opposition to Knoch (2009). Monteiro (2019) investigated the proficiency construct of intercultural radiotelephony communications in aviation and specified the communicative demands of pilots and ATCOs within a construct framework. She points out that pilots and ATCOs perceive that intercultural factors can impact the safety of flights. The results of her study indicate that the ICAO policy does not include important components of the construct. Monteiro (2022) highlights that

Effective RT communications require competencies not addressed in prevailing models of communicative competence. They do require specific purpose language ability and background knowledge (AE), the need to speak English as a lingua franca and to adjust to the communicative needs at hand (ELF), to accommodate and negotiate sociocultural differences (ICA), and to solve misunderstandings between members of different cultures, while at the same time sharing responsibility for successful communication (IC). And most importantly, this applies to both first language (L1) speakers of English, and those who speak English as a second (L2) or additional language. (Monteiro 2022: 239)

As seen in this brief literature review, there has been a quite prolonged debate on the appropriacy of the ICAO policy. As Knoch, Deygers, and Khamboonruang (2021) point out, the rating scale “includes features of construct irrelevance and construct under-representation of the TLU domain” (Knoch, Deygers, and Khamboonruang 2021: 618). We can then conclude that “the ICAO policy has not met its intended goals, and these seem unlikely to be met in the future unless the policy and its underlying construct are modified” (Kim 2013: 108).
3. METHODS

3.1. What Document was Selected?

Merriam (1988) argues that “documents of all types can help the researcher uncover meaning, develop understanding, and discover insights relevant to the research problem” (Merriam 1988: 118). Policy documents carry valuable information that can offer insights as to how such policies should be implemented.

The source of data for this study was the ICAO Rating Scale, included in the Annex 1. When planning this project, the intention was to analyse not only the rating scale, but the whole ICAO policy, including the Holistic Descriptors and the rest of ICAO Annex 1 (ICAO 2022a), the second edition of the ICAO DOC 9835 (ICAO 2010), the ICAO Circular 318 (ICAO 2009), Language Testing Criteria for Global Harmonization (ICAO 2009), and the ICAO test design guidelines (ICAEA n.d.). However, the ICAO Annex 1 is the only document that has the Standard and Recommended Practices (SARPs), or, in other words, the requirements that Contracting States need to comply with, a decision was made to only analyse this fundamental document. The manual, the circular, and the test design guidelines contain guidelines, which are very relevant but unfortunately not mandatory for Contracting States to follow. However, when discussing the results, I occasionally mention DOC 9835. However, the most relevant piece of the policy to define the construct is the rating scale, which every test instrument developed to assess pilots’ and ATCO’s ability to speak and understand the language used in radiotelephony communications must employ.

3.2. What Method was Applied?

The method applied in this qualitative study was document content analysis. Bowen (2009) defines document analysis as “a systematic procedure for reviewing or evaluating documents” (p. 27), which “entails finding, selecting, appraising (making sense of), and synthesising data contained in documents” (p. 28). This procedure is recommended by Cardno (2018) to analyse the organization and content of educational policy documents. As Cardno (2018) explains, “as a research tool, policy document analysis is a method for investigating the nature of a policy document in order to look at both what lies behind it and within it” (Cardno 2018: 625). In spite of the traditional quantitative nature of content analysis (Merriam, 1998), the focus of the present analysis was not on the quantitative aspects of the policy content, such as frequencies, but on its qualitative nature, such as “the presence or absence of certain content characteristic” (George 2009: 145).

Bowen (2009) explains that document analysis is an iterative process which includes superficial skimming, careful reading, and interpretation. This document analysis was done through a careful reading of the document, and involved focused re-reading and review of the documentary data. Thus, an “interview technique”, as described by O’Leary (2021), was conducted to interrogate the rating scale as if the text was being interviewed. As O’Leary (2021) points out:

In ‘interviewing’ your documents, you are, in a sense, treating each document as a respondent who can provide you with information relevant to your enquiry. The questions you ask will be dependent on the nature of your enquiry and on the document type. As with an interview, you will need to determine what it is you want to know, and whether your document can provide you with the answers. You then need to ‘ask’ each question and highlight the passages in the document that provide the answer. (O’Leary 2021: 200)
Having RQ1 in mind, the following questions were asked towards the document:
Q1) **What are the common elements in the comprehension descriptor levels?**
Q2) **What elements are not recurrent among the comprehension descriptor levels?**
Q3) **How do the comprehension descriptors differentiate the different levels?**
Q4) **Do the comprehension descriptors make a difference between a test to assess interactive listening and a test to assess listening in isolation?**

While highlighting the common elements in the comprehension descriptors and comparing the differences between the levels, some other questions arose:
Q5) **What may be considered a common topic?**
Q6) **What may be considered a concrete topic?**
Q7) **What may be considered a work-related topic?**
Q8) **What accents or varieties may be considered sufficiently intelligible for an international community of users?**
Q9) **What can be considered a linguistic complication?**
Q10) **What can be considered a situational complication?**
Q11) **What is an unexpected turn of events?**
Q12) **What is meant by nearly all contexts?**

### 4. RESULTS AND DISCUSSION

#### 4.1. What is the Listening in Isolation Construct Explicitly Stated in the ICAO Language Proficiency Rating Scale?

The common element in the comprehension descriptors from level 3 to level 5 is that they talk about comprehension in common, concrete, and work-related topics. Another common element is that the speaker might be confronted with a linguistic or situational complication or an unexpected turn of events. Table 1 shows the differences between comprehension in these three levels. We can see that comprehension in common, concrete, and work-related topics seems to be easier, as comprehension of test takers who will not even pass the test (level 3) is often accurate and comprehension of level 4 test takers is mostly accurate (although it allows a few misunderstandings). When the speaker is confronted with a linguistic or situational complication or an unexpected turn of events, the descriptors seem to be harsher on test takers. For test takers to pass the test (get a level 4), although it might take them some time or the use of clarification strategies, they seem to need to understand the communication, as only level 3 test takers may fail to understand.

<table>
<thead>
<tr>
<th>Level</th>
<th>Comprehension in common, concrete, and work-related topics</th>
<th>Comprehension when the speaker is confronted with a linguistic or situational complication or an unexpected turn of events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extended 5</td>
<td>Accurate</td>
<td>Mostly accurate</td>
</tr>
<tr>
<td>Operational 4</td>
<td>Mostly accurate</td>
<td>May be slower or require clarification strategies.</td>
</tr>
<tr>
<td>Pre-operational 3</td>
<td>Often accurate</td>
<td>May fail to understand</td>
</tr>
</tbody>
</table>
We can see that the difference between comprehension in levels 3, 4, and 5 is how accurate the comprehension on common, concrete, and work-related topics of the test taker is (level 5, accurate, level 4, mostly accurate, and level 3, often accurate). For levels 3 and 4, the scale includes an observation that for comprehension in common, concrete and work-related topics to be either mostly accurate (level 4) or often accurate (level 3), the accent or variety used should be sufficiently intelligible for an international community of users. The absence of this observation in the level 5 descriptors may imply that, at this level, comprehension must be accurate even when the accent or variety used is not sufficiently intelligible for an international community of users.

As mentioned, during the analysis, questions 5 to 12 above were risen. Some of these questions had been raised in previous work. Douglas (2004), for example, asked two important questions: “Does the phrase ‘common, concrete, and work-related topics’ refer to three topic categories or one?” and “how is the intelligibility of the various English dialects and accents in use internationally to be determined?” (Douglas 2004: 250). Knoch (2009) suggested that this reference to intelligibility for an international community of users could be deleted from the rating scale because it would be problematic to select speakers that would satisfy this requirement. Garcia (2015) recommends that this issue should be further researched.

The descriptors for level 6, on the other hand, instead of describing comprehension in the two mentioned scenarios, talk about comprehension in nearly all contexts. What is meant by “nearly all contexts”? Is it more than comprehension in both mentioned scenarios? Apparently, yes, as the word all implies. But why nearly all contexts? What contexts are not included?

DOC 9835 explains that

Work-related context can accommodate different interpretations. A narrow interpretation would aim to closely replicate radiotelephony communications, including the extent of plain language needed in unusual, unexpected or emergency situations. A broad interpretation of the holistic descriptors and Rating Scale would aim to elicit plain language on various topics that are related to radiotelephony communications or aviation operations, without replicating radiotelephony communications specifically. … Both interpretations are valid. (ICAO 2010: 6.2.8.9)

Not having the “right” interpretation to be followed determined by ICAO gives too much flexibility for test developers to define the construct they want to measure. The fact that the DOC 9835 allows for different interpretations contributes to the significant differences in test design which have led to uncertainties in relation to what tests measures, what results mean, and overall quality of tests worldwide (ICAO 2022b). Unfortunately, although this issue of having a possibility for either a broad or a narrow interpretation was frequently brought to discussion in the ICAO EG-LPRs/03 meetings mentioned in Note 3, this will remain an unresolved problem for some time to come.

Furthermore, some elements in the comprehension descriptors are only mentioned in one of the levels. These are: the reference to the ability “to comprehend a range of speech varieties … or registers”, which are only included in level 5, and the ability to comprehend linguistic and cultural subtleties, which are only mentioned in level 6. The issues in the ICAO rating scale reported here, such as the terminology problems, inconsistencies, and lack of definition of concepts do not only happen in this scale.
Alderson et al. (2004) has also listed similar problems regarding the Common European Framework. These problems make it difficult to fully understand the construct to be measured, but, within our limitations, we can conclude that a test that aims to assess pilots and ATCOs listening comprehension in isolation must include the abilities listed in Table 2. The comprehension descriptors seem to apply to the assessment of both listening in isolation and interactive listening.

Table 2 Listening construct represented in the comprehension descriptors

<table>
<thead>
<tr>
<th>Item</th>
<th>Construct</th>
<th>Applicability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Comprehension on common, concrete, and work-related topics</td>
<td>Levels 3, 4, 5</td>
</tr>
<tr>
<td>2</td>
<td>Comprehension when the speaker is confronted with a linguistic</td>
<td>Levels 3, 4, 5</td>
</tr>
<tr>
<td></td>
<td>or situational complication or an unexpected turn of events.</td>
<td></td>
</tr>
<tr>
<td>2.1</td>
<td>In assessing 1, include accents or varieties sufficiently</td>
<td>Levels 3, 4</td>
</tr>
<tr>
<td></td>
<td>intelligible for an international community of users.</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Ability to ask for clarification when comprehension fails</td>
<td>Level 4</td>
</tr>
<tr>
<td>4</td>
<td>Comprehension of a range of speech varieties (dialect and/or accent) or</td>
<td>Level 5</td>
</tr>
<tr>
<td></td>
<td>registers.</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Comprehension in nearly all contexts</td>
<td>Level 6</td>
</tr>
<tr>
<td>6</td>
<td>Comprehension of linguistic and cultural subtleties</td>
<td>Level 6</td>
</tr>
</tbody>
</table>

Test developers need to have clear definitions for the concepts that are not defined. An analysis of the ICAO documents that include guidelines for the implementation of the ICAO LPRs, such as the ones mentioned in 3.1, may help to further understand ICAO’s intentions. The second edition of DOC 9835 includes an explanation of the rating scale descriptors from level 3 to 6 which may be useful in this process.

4.2. How May the Descriptors for the Criteria other than Comprehension Help to Define the Construct of a Test to Assess Listening in Isolation?

Questions 5 to 12 were addressed towards the rating scale in order to investigate how the descriptors for the other criteria could help to define the construct of a test to assess listening in isolation. The rating scale does not provide a clear answer to questions 5, 6, and 7, as it does not explain what can be considered a common, concrete, and work-related topic. However, these terms also appear in the descriptors for vocabulary from level 3 to 5. In the vocabulary descriptors, the frequency in which vocabulary range and accuracy are sufficient to communicate on common, concrete, and work-related topics should be evaluated. According to Knoch (2009), “reference to ‘common, concrete and work-related topics’ is not clear to stakeholders as the whole test should be in the aviation domain” (Knoch 2009: 43). DOC 9835 states that

Context is an important consideration in communications, and an individual’s language proficiency may vary in different contexts. This holistic descriptor limits the domain of the communicative requirements to work-related topics; that is, air traffic controllers and pilots are expected to be able to communicate about issues in their field of professional practice. Language proficiency should not be limited to standardized phraseology and should range across a relatively broad area of work-related communicative domains. (ICAO 2010: 4.5.3)
Very high lexical familiarity seems to be essential for good comprehension (Bonk 2000). According to Rost and Brown (2022), word recognition is the basis of spoken language comprehension, and lexical knowledge seems to be, in second language listening, “the most significant variable contributing to listening proficiency” (Rost and Brown 2022: 241). Thus, test takers’ lexical knowledge should also be assessed in a listening test. Elements of the vocabulary descriptors, such as knowledge of vocabulary used on common, concrete, and work-related topics, of idiomatic vocabulary, of vocabulary used on a wide variety of familiar and unfamiliar topics, and of nuanced vocabulary can inform the definition of the construct of a test to assess listening in isolation. It is important to point out that the purpose of the present study is to try to understand the construct represented in the ICAO rating scale, not to evaluate or criticize it. However, it is good to mention that criticism (e.g. Garcia 2015; Knoch 2009; Pfeiffer 2009) has been made in relation to the fact that the rating scale descriptors include idiomatic and nuanced vocabulary since communications between pilots and ATCOs should always be “clear, concise, and unambiguous”, even when using plain language (ICAO 2010: 4.3.4). Knoch (2009), for example, argues that “any references to idiomatic language should be deleted as this is not appropriate in the TLU domain” (Knoch 2009: 43). I have also strongly argued that “idiomatic vocabulary should never have been included in the rating scale” (Garcia 2015: 22). Today I think differently because unfortunately real-life communications are not always according to standards. As a matter of fact, ICAO acknowledges, as mentioned earlier, that 70% of the radiotelephony speech acts do not comply with the recognized standards (ICAO 2010). Thus, I argue that if idiomatic expressions are used in real-life radiotelephony communications, as the findings of Prinzo, A. Hendrix, and R. Hendrix (2009) indicate, they should be included in a test that intends to reflect the language used in real-life. Although I believe pilots and ATCOs should have the ability to comprehend idiomatic vocabulary assessed, they should be encouraged not to use them when speaking. Also, the vocabulary descriptors for level 6 may help to understand what the comprehension descriptors for level 6 mean by “a wide variety of familiar and unfamiliar topics.”

The descriptors for pronunciation, which include both segmental and suprasegmental features of pronunciation, may help to address question 8. From level 2 to 6, “pronunciation, stress, rhythm, and intonation” might be “influenced by the first language or regional variation”. However, the pronunciation of a level 2 test taker is heavily influenced, while the pronunciation of a level 6 might be influenced or not. The main difference between the levels is the frequency in which pronunciation interferes with ease of understanding. This frequency may “usually” (level 2), “frequently” (level 3), “only sometimes” (level 4), “rarely” (level 5), or “almost never” (level 6) interfere with ease of understanding. The accents and varieties that may be considered sufficiently intelligible for an international community of users, as mentioned in the comprehension descriptors for levels 3 and 4, might be the ones whose pronunciation either rarely (level 5 in pronunciation) or almost never (level 6 in pronunciation) interferes with ease of understanding. Thus, a recording used in a listening test recorded by a pilot or controller who was awarded level 5 or 6 in pronunciation might be considered an accent or variety that is sufficiently intelligible for an international community of users, whereas recordings which were recorded by pilots or controllers who were awarded level 4 in pronunciation might be considered within a range of speech varieties, which is mentioned in the comprehension descriptors for level 5. Test developers may even consider having pilots or controllers who were awarded 3 or less in pronunciation to make the recordings
for the listening test, since ICAO’s report on the implementation of the LPRs showed that 32.25% of member states had not provided information about their implementation status (ICAO 2013). However, it is important to point out that “although strength of foreign accent is indeed correlated with comprehensibility and intelligibility, a strong foreign accent does not necessarily cause L2 speech to be low in comprehensibility or intelligibility” (Munro and Derwing 1999: 305).

Moving on, what could be considered a linguistic complication (question 5 from Table 2)? As linguistic is a broad term, many linguistic factors may be considered a linguistic complication. The interference of pronunciation, stress, rhythm, or intonation with ease of understanding may be considered linguistic complications. As Rost and Brown (2022) argue, “unexpected speaker accents, an unfamiliar phonotactic pattern, rhythm and intonation systems, length of input, number of speakers, rapid speech rate and lack of pauses, and connected speech phenomena (reductions and assimilations)” may pose challenge to linguistic processing (Rost and Brown 2022: 249). Listening to unfamiliar or uncommon vocabulary may also cause a linguistic complication. Even factors related to the fluency descriptors may add a complication, as “lack of ‘orality’ features (such as pauses and redundancy)” may also pose challenges. Furthermore, lack of knowledge of syntax may also be considered a linguistic complication. Rost and Brown (2022) argue that for a listener to have a detailed comprehension of a message, “a thorough syntactic processing needs to take place” (Rost and Brown 2022: 242). The descriptors for structure talk about basic and complex grammatical structures. Although there has been a discussion whether the scale’s reference to complex structures reflects the real-world of radiotelephony communications (e.g., Prado 2015), there is no doubt that knowledge of grammatical structure plays an important role in comprehension. Rost and Brown (2022) list complexity of grammatical structures as one factor that poses challenges to second language listeners.

Now, what is a situational complication (question 10)? According to DOC 9835, “it is during complications in aviation that communications become most crucial, with a greater reliance upon plain language” (ICAO 2010: 4.6.6). The descriptors for vocabulary and interactions may shed some light on this issue. First, when the descriptors for vocabulary mention common topics, they imply topics might also be uncommon topics. They also talk about unfamiliar topics. Differently from uncommon topics, which may be understood as topics that do not happen frequently in radiotelephony communications, unfamiliar topics are the ones that pilots and ATCOs were not familiar with. When topics are uncommon or unfamiliar, the situation will likely be more difficult to solve. The level 4 descriptors for interactions talk about the ability to deal adequately with apparent misunderstanding. The occurrence of a misunderstanding could also be considered a situational complication. The topic of question 7 (“an unexpected turn of events?”) may also be considered a possible situational complication, as it describes a situation in which events do not occur according to what is expected. DOC 9835 points out that

One of the more challenging events in all communications, including those involving the use of a second language, is when the unexpected happens. Human Factors experts have emphasized the threat of letting our expectations hinder our interpretation of reality. Sometimes, a complication or an unexpected event can lead to a communication breakdown. (ICAO 2010: 4.5.3)

The opposite of “an unexpected turn of events”, according to the level 3 descriptors for structure and interactions, seems to be “predictable situations.” Moreover, if a
response is not immediate, appropriate, or informative (as the descriptors for interactions from level 3 to 5 describe), a situational complication may arise.

Table 3 shows a summary of elements test developers may consider when developing a listening test to assess pilots and ATCOs’ listening in isolation, which was based on the present analysis of how other descriptors other than comprehension may help to inform the definition of the construct.

Table 3 Elements test developers may consider when developing a listening test to assess pilots and ATCOs’ listening in isolation

<table>
<thead>
<tr>
<th>Include pronunciation, stress, rhythm, and intonation both influenced and not influenced by the first language or regional variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>The accents and varieties that may be considered sufficiently intelligible for an international community of users, as mentioned in the comprehension descriptors for levels 3 and 4, might be the ones whose pronunciation either rarely (level 5 in pronunciation) or almost never (level 6 in pronunciation) interferes with ease of understanding</td>
</tr>
<tr>
<td>Include a range of speech varieties (one parameter might be pilots and ATCOs who were awarded level 4 or below in pronunciation)</td>
</tr>
<tr>
<td>Include the assessment of the ability to understand vocabulary used on common, concrete, and work-related topics, idiomatic vocabulary, vocabulary used on a wide variety of familiar and unfamiliar topics, nuanced vocabulary</td>
</tr>
<tr>
<td>Comprehension of a wide variety of familiar and unfamiliar topics may help to understand what the comprehension descriptors for level 6 mention as comprehension “in nearly all contexts”</td>
</tr>
<tr>
<td>A linguistic complication might be caused by:</td>
</tr>
<tr>
<td>▪ Interference of pronunciation, stress, rhythm, or intonation, a linguistic complication on ease of understanding</td>
</tr>
<tr>
<td>▪ The use of complex structures</td>
</tr>
<tr>
<td>▪ The presence of uncommon or unfamiliar vocabulary</td>
</tr>
<tr>
<td>▪ Inappropriate phrasing and pausing, slowness in producing language, use of too many fillers, or use of inappropriate discourse markers or connectors</td>
</tr>
<tr>
<td>An unexpected turn of events is one situational complication, and might be caused by a misunderstanding, or by a response which is not immediate, appropriate, or informative, among other possibilities.</td>
</tr>
<tr>
<td>Include predictable situations (to contrast with unexpected situations)</td>
</tr>
</tbody>
</table>

5. Conclusion

Investigating the construct of the ICAO rating scale is only one piece of the puzzle to define the construct of a test to assess listening in isolation. Investigating the policy is one of the five different areas test developers of Language Assessments for Professional Purposes (LAPPs) should consider when developing a test, as Knoch and Macqueen (2020) recommend. They should also analyse the needs and motivations of test takers, test requirements needs, the availability of resources for test development, administration, and validation, and, most importantly, the characteristics of the TLU domain. Furthermore, the present study only analysed part of the policy, as the focus was on the rating scale.
Knoch and Macqueen (2020) suggest including in the policy analysis, an analysis of the policy environment and of the professional registration environment.

As Upshur and Turner (1995) point out, “in general, … rating scales present major problems of reliability and validity” (Upshur and Turner 1995: 5). Not differently, the ICAO rating scale is problematic and revision should be considered by ICAO (Aragão 2018; Garcia 2015; Knoch 2009; Pfeiffer 2009). The fact that the rating scale was developed to assess speaking makes it difficult to be used to assess listening (Knoch 2009). As Garcia (2015) suggests, the development of a specific rating scale for the assessment of comprehension seems to be necessary. Future studies of the ICAO rating scale can inform its revision and contribute to the development of a more valid and reliable scale to assess speaking and of a new rating scale to assess listening. Researchers and test developers could look into other dimensions of construct as described by Knoch and Macqueen (2020): the perceived construct (e.g., how stakeholders understand the construct in the rating scale) and the operationalized construct (what is actually being assessed by tests). Further analysis of other aspects of the ICAO policy may offer insights into understanding problems related to the implementation of the ICAO LPRs. For example, researchers and test developers could also look at aspects related to the construction of the policy, its context, and impact, including its strengths and concerns. Moreover, researchers and test developers could also investigate the values embodied in the policy (see Slohamy 2001). Furthermore, they could conduct an evaluation of the policy using the policy evaluation framework provided by Knoch and Macqueen (2020).

The purpose of this study was to improve the understanding of the ICAO policy in order to inform the development of tests that aim to assess pilots and ATCOs’ listening in isolation fairly, as required by the ICAO test design guidelines (ICAEA n.d.). Having a common understanding of the construct may contribute to reduce the variation of quality of tests worldwide. To conclude, test developers need to keep in mind that

In such a ‘high-stakes’ environment, language testing needs to be accountable to the stakeholders in the aviation industry. State regulators, managers of airline and air traffic management service providers, trainers, pilots and controllers, and ultimately, the flying public, need to be able to trust global language assessments and to have confidence that licensed operations personnel are competent communicators in the English language. As those who are obliged to learn and use English on the frequency, pilots and controllers deserve to have their language proficiency assessed fairly, and to know that their counterparts around the world have been assessed according to the same standard. (Emery 2007: 1)

REFERENCES


Investigating the Construct of Aeronautical English Listening Testing


References


Garcia, Angela C. M. *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*. Unpublished Master’s Thesis. Lancaster University, 2015.


Knoch, Ute, Bart Deygers, and Apichat Khamboonruang. “Revisiting Rating Scale Development for Rater-Mediated Language Performance Assessments: Modelling


Chapter 5: Manuscript 4 – Contexts and constructs: Implications for the testing of listening in pilots’ communication with air traffic controllers

The manuscript in this chapter consists of an article published in *The ESPecialist* journal in 2020. This journal is one of the earliest applied linguistics journals in Brazil, and is published by the Pontifical Catholic University of São Paulo. This publication resulted from the pilot study I conducted in the first year of my doctoral program at Carleton University under the supervision of Dr. Janna Fox and Dr. Natasha Artemeva. The purpose of this empirical study was to investigate the perceptions of pilots and other key stakeholders on the listening tasks performed by pilots in their communications with ATCOs in order to inform the definition of the listening construct of a test that aims to assess pilots’ listening comprehension.

Alternatively, the article can be read through the following hyperlink:


The following corrections/suggestions of improvement should be acknowledged:

<table>
<thead>
<tr>
<th>Page number/line</th>
<th>From</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>11/2</td>
<td>Five closed-ended opinion questions</td>
<td>Nine closed-ended opinion questions</td>
</tr>
</tbody>
</table>
**Contexts and constructs: Implications for the testing of listening in pilots’ communication with air traffic controllers**

Angela C. M., GARCIA¹
Janna, FOX²

**ABSTRACT**
The International Civil Aviation Organization (ICAO) published the Language Proficiency Requirements for pilots and air traffic controllers (ATCOs) in 2003. Research has shown that there is still a lack of clarity regarding what needs to be assessed in terms of the English used by pilots and ATCOs in radiotelephony (DOUGLAS, 2014; EMERY, 2014; KIM; BILLINGTON, 2016; KIM; ELDER, 2015; MONTEIRO, 2019; READ; KNOCH, 2009). The purpose of the present study was to investigate the nature of the listening tasks performed by pilots as an essential step in test development and validation. This explanatory sequential mixed-methods study (CRESWELL, 2015), elicited questionnaire responses from 156 pilots (Phase 1) and subsequently interviewed six aeronautical English experts (Phase 2) to better understand the characteristics of the listening tasks that pilots undertake during radiotelephony communications. Quantitative and qualitative data were analysed, and findings were merged. They provide information that may usefully inform the development of the listening test construct and the test specifications.

**Keywords:** ESP Assessment, Aeronautical English, Listening Comprehension, Testing Listening, ICAO

**RESUMO**
A Organização Internacional de Aviação Civil (OACI) publicou os Requisitos de Proficiência Lingüística para Pilotos e Controladores de Tráfego Aéreo (ATCOs) em 2003. Pesquisas mostraram que ainda há falta de clareza em relação ao que precisa ser avaliado no que se refere ao inglês usado pelos pilotos e ATCOs na radiotelefoneia (DOUGLAS, 2014; EMERY, 2014; KIM; BILLINGTON, 2016; KIM; ELDER, 2015; MONTEIRO, 2019; READ; KNOCH, 2009). O objetivo deste estudo foi investigar a natureza das tarefas de compreensão auditiva realizadas pelos pilotos, como uma etapa essencial no desenvolvimento e validação de testes. Este estudo explanatório sequencial de métodos mistos (CRESWELL, 2015), suscitou respostas de 156 pilotos a um questionário (Fase 1) e posteriormente entrevistou seis especialistas em inglês aeronáutico (Fase 2) para um melhor entendimento das características das tarefas de compreensão auditiva que os pilotos realizam durante as comunicações de radiotelefoneia. Dados quantitativos e qualitativos foram analisados e os resultados foram integrados. Eles fornecem informações que podem ser úteis para o desenvolvimento do contrato de teste de compreensão auditiva e das especificações de teste para pilotos.

**Palavras-chave:** Avaliação De Inglês Para Fins Específicos, Inglês Aeronáutico, Compreensão Auditiva, Avaliação da Compreensão, OACI

* This article is part of a PhD’s research conducted by the first author at the School of Linguistics and Language Studies of Carleton University, under the supervision of the second author.
1 Carleton University, Ontario, Canada. PhD student in Applied Linguistics & Discourse Studies; School of Linguistics and Language Studies; ORCID: [https://orcid.org/0000-0003-3018-2370](https://orcid.org/0000-0003-3018-2370); angelagarcia@e-mail.carleton.ca
2 Carleton University, Ontario, Canada. Professor Emeritus at the School of Linguistics and Language Studies; ORCID: [https://orcid.org/0000-0003-4528-4513](https://orcid.org/0000-0003-4528-4513); jannafox@eunet.carleton.ca
1. **Introduction**

Since 5 March 2008 (later extended to 2011), pilots and air traffic controllers (ATCOs) have been required to demonstrate their ability to speak and understand the language used for radiotelephony (RT) communications in accordance with the International Civil Aviation Organization’s (ICAO) Language Proficiency Requirements (LPRs). Although the adoption of the ICAO LPRs was a tremendous advance, several issues have been identified with the policy (e.g., DOUGLAS, 2014; EMERY, 2014; KIM; BILLINGTON, 2016; KIM; ELDER, 2015; MONTEIRO, 2019; READ; KNOCH, 2009). The main issue is that the test construct is underrepresented. Emery (2014), for example, argued that “the ICAO guidance … is of little practical use in the definition of the test construct and the development of test specifications” (EMERY, 2014, p. 206), and Douglas (2004) pointed out that “we need much more information about the nature of language use in aviation contexts outside the prescribed phraseologies” (DOUGLAS, 2004, p. 250). The term construct is defined by Bachman (2007) simply as “what it is that we want to assess” (BACHMAN, 2007, p. 42). He argues that understanding the roles of abilities, contexts and their interactions has been a central problem in the field of language assessment, especially because the way these roles are viewed has a significant impact on how a construct is defined. Construct underrepresentation occurs when a test “is too narrow and fails to include important dimensions or facets of focal constructs” (MESSICK, 1996, p. 4), and this represents a threat to validity.

One of the problems regarding the ICAO policy lies in the fact that listening comprehension represents only one of six criteria included in the ICAO rating scale (GARCIA, 2015). The ICAO rating scale consists of descriptors for the assessment of six different categories or skills: pronunciation, structure, vocabulary, fluency, comprehension, and interactions, that range from Level 1 (Pre-elementary) to Level 6 (Expert). Level 4 (operational) is the minimum level required for international operations. However, in the second edition of ICAO’s DOC 9835 (2010), *Manual on the implementation of ICAO Language Proficiency Requirements*, ICAO included the information that “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). Garcia (2015) interviewed expert ICAO language test developers and raters to discuss the rating scale descriptors. One of the participants remarked that he did not think that “the listening construct in the case of RT communications is adequately defined or captured by the criteria as they stand” (GARCIA, 2015, p. 93). He argued that not only was comprehension inadequately measured, but also its importance was undermined by the rating scale:

[http://revistas.pucsp.br/esp](http://revistas.pucsp.br/esp)  
DOI: [https://doi.org/10.23925/2318-7115.2020v41i4a4](https://doi.org/10.23925/2318-7115.2020v41i4a4)
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. (GARCIA, 2015, p. 38)

As a result, the author concluded that “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” (GARCIA, 2015, p. 62). As pointed out by Kim and Elder (2015), “a single piece of unclarified information could have disastrous results in air traffic control” (KIM; ELDER, 2015, p. 133). Thus, it is extremely important to investigate the nature of the listening comprehension skills of pilots in their communications with the ATCOs. Only with a clear, empirically sound understanding will a test construct adequately represent the features of the target language use (TLU) tasks (BACHMAN; PALMER, 1996). As Field (2019) argued, “when designing a test of one of the language skills, we need extensive information about the nature of the phenomenon we are dealing with. This enables us to shape the material and tasks that we devise so as to ensure that they truly measure competence in the skill” (FIELD, 2019, p. 7). Therefore, the goal of the present study is to identify the characteristics of pilots’ listening tasks in the context of RT communications in order to better define the test construct. It aims to contribute to the creation of knowledge regarding the construct of the ICAO LPRs for pilots and ATCOs, and also to contribute to a future revision of the rating scale descriptors. The findings from this study may be useful in informing the development of ICAO language proficiency tests, as well as supporting test developers in the design of listening tests and tasks that are more appropriate to the TLU domain.

As background to the study, in the section which follows below, a review of relevant literature in the fields of language assessment and aeronautical English is provided. This is followed by the specific research questions that guided the study, the methods, and results and discussion. Concluding remarks include the limitations of the study and provide some suggestions for further research.

2. Literature review

Listening is complex, especially in a second or foreign language. Assessing listening comprehension is challenging, particularly in the case of assessment of Languages for Specific Purposes (LSP), which can be understood as:

http://revistas.pucsp.br/esp

DOI: https://doi.org/10.23925/2318-7115.2020v41i4a4
[...] one in which test content and methods are derived from an analysis of a specific purpose target language use situation, so that test tasks and content are authentically representative of tasks in the target situation, allowing for an interaction between the test taker’s language ability and specific purpose content knowledge, on the one hand, and the test tasks on the other. Such a test allows us to make inferences about a test taker’s capacity to use language in the specific purpose domain. (DOUGLAS, 2010, p. 19)

This literature review is divided into three subsections: 1) a selective overview of research on the assessment of listening; 2) the nature of listening in the context of pilots/ATCOs RT communications; and 3) the assessment of pilots’ listening ability.

2.1 The assessment of listening

As Buck (2001) pointed out, if we want to assess listening, it is important to first understand the process of listening, which is very complex. Listening can be understood as the recognition and comprehension of spoken language. Listening, similarly to reading, but differently from speaking and writing, is an internal process, i.e. a receptive skill, as it is not observable (WAGNER, 2014). As Field (2019) pointed out, “the listening process takes place in the mind of the listener” (FIELD, 2019, p. 1). At the same time, listening is an active process, as listeners are not empty containers into which the information is poured (LYNCH; MENDELSOHN, 2010). Buck (2001) found that listeners interpret the information they receive according to their linguistic knowledge (e.g., phonology, lexis, syntax, semantics and discourse structure) and their non-linguistic knowledge (e.g., their topical knowledge, their contextual knowledge, their individual characteristics, and their background knowledge). As he pointed out, “when we listen, we use our background knowledge of the world to set up expectations, and then use those expectations to help us comprehend what we hear” (BUCK, 2001, p. 8). Those expectations might be different among listeners, along with their motives for listening. As Buck (2001) explained, different people have different interests and different needs. Furthermore, listening, as discussed by Lynch and Mendelsohn (2010), “is not merely an auditory version of reading” (LYNCH; MENDELSOHN, 2010, p.180). It disappears as quickly as it appears, and patterns of rhythm, sound, stress, and intonation are key to understanding it. Moreover, it has the presence of natural fast speech features and the frequent need to interact immediately. In other words, listening is “an interactive process” (BRINDLEY, 1998, p. 172).

The assessment of listening, as Wagner (2014) noted, “has historically been somewhat neglected and even overlooked in the language literature” (WAGNER, 2014, p.1). Test developers and item writers face some challenges when designing listening tests and creating tasks. It is difficult to describe and to

http://revistas.pucsp.br/esp

DOI: https://doi.org/10.23925/2318-7115.2020v41i4a4
assess “an invisible cognitive operation” (BRINDLEY, 1998, p. 171), such as listening. For example, it is challenging for test writers to create tasks that will not only elicit the use of receptive skills, but also result in a situation where the test-taker will have the opportunity to demonstrate that they can use their skills successfully (HUGHES, 2003).

Test developers face some other challenges when writing test specifications, such as, deciding whether or not the candidates will be allowed to listen to the text more than once (TAYLOR; GERANPAYEH, 2011), issues related to task authenticity (ALDERSON et al., 1995; BRINDLEY, 1998; HUGHES, 2003; LYNCH; MENDELSOHN, 2010; WAGNER, 2014), to memory (BUCK, 2001; WU, 1998), and to the complexities of the cognitive processes involved in listening (TAYLOR; GERANPAYEH, 2011). Therefore, test developers need to think carefully about the purpose of the test and the specific target language situation in order to make informed decisions.

2.2 The nature of listening in pilots/ATCOs RT communications

RT communications between pilots and ATCOs are very different from normal every-day conversations. Pilots and ATCOs use standardized phraseology, a formulaic code made up of specific words and phrases, in order to communicate via the radiotelephone. Whenever phraseology does not suffice, they must resort to plain language, which is defined as “the spontaneous, creative and non-coded use of a given natural language” (ICAO, 2010, p. x). Listening in pilot/ATCO communication consists of a two-way listening mode as it involves both listening and speaking. As pilots and ATCOs talk on the radio, they must take turns. There is no overlapping, which is very common in other conversational contexts. This kind of communication “represents a very specialized and socially significant form of discourse” (READ; KNOCH, 2009, p. 21.3).

In a typical transmission, the ATCO sends a message, the pilot listens to it, repeats it to the ATCO (readback), and the ATCO listens to it and confirms it (hearback). Research has shown that visual support helps listeners to understand spoken texts (BUCK, 2001). However, pilot/ATCO RT communications are voice-only interactions, with no facial references. For this reason, it is very important for pilots and ATCOs to be collaborative listeners. Collaborative listening comprises “making appropriate requests for clarification, back-channelling, making responses to interactional language, or taking responsibility for organising turn-taking” (BUCK, 2001, p. 12).

Both native English speakers (NES) and non-native English speakers (NNES) make mistakes when trying to understand an ATCO’s transmission. However, NNES pilots seem to be more challenged
by the English used in RT communications (WU et al., 2019). Wu et al. found that, NNES pilots “make more overall errors in their transmissions than native English sounding pilots” (WU et al., 2019, p. 9).

A number of factors may make it difficult for pilots to understand ATCOs. Emotional stress or increased mental workload may impact language production, as people have the tendency, under these conditions, to talk louder and faster (PRINZOl; BRITTON, 1993, p. 16). Moreover, Estival and Molesworth’s (2016) study showed that increased pilot workload has a negative impact on the accuracy of pilots’ readbacks.

Faster speech rate is another factor that might increase communication difficulty for pilots and ATCOs. Prinzo and Britton (1993) claimed that, even under normal conditions, ATCOs tend to speak too fast, and that pilots also talk faster during heavy workload. Research has indicated that “the faster the speech, the more difficult it is to comprehend” (BUCK, 2001, p. 38). Additionally, it is important to point out that there might be a relationship between how much speech rate interferes with understanding speech and the test taker’s proficiency level, although, according to Révézé and Brunfaut’s (2013) study, an increased speech rate did not have a significant effect on advanced level test takers’ results in listening tasks. Experience seems to be another important factor. Estival and Molesworth’s (2016) study showed that a faster speech rate, if combined with lower experience, might still be a problem for more advanced speakers. Their results showed that speech rate had a greater impact on communication accuracy of low qualified NES pilots (private pilot licence or lower) and English as a second language pilots than on high qualified NES (commercial pilot license or higher), as the former “were adversely affected by faster ATC speech rate” (ESTIVAL; MOLESWORTH, 2016, p. 234).

Furthermore, accent plays an important role in relation to how easily a pilot understands an ATC communication. Pilots and ATCOs are exposed to a wide variety of unfamiliar accents. In many interactions, both participants are NNES and come from different cultural backgrounds. As Kim and Billington (2016) pointed out, “given the likelihood that both pilots and controllers will encounter English in a variety of accents and proficiency levels, their ability to effectively comprehend a range of accents is

---

1 According to Wu et al.’s (2019) study, “pilots were classified as ‘native English sounding’ if no foreign accent could be detected and if they were flying with an airline registered in a country where the official language is English (e.g., Australia, New Zealand, the United Kingdom, or the United States of America). Pilots were classified as ‘accented English’ if a non-native English accent could be detected and if the aircraft was registered in a country where English is not one of the official languages (e.g., Japan, Korea, China, Chile). Pilots with an English sounding accent, but who were on an aircraft that was registered in a country where English is not the official language or is one of several official languages (e.g., Hong Kong, Singapore, Malaysia, Fiji) were excluded, as it was less certain whether their native language was English” (WU et al., 2019, p. 3).

http://revistas.pucsp.br/esp

DOI: https://doi.org/10.23925/2318-7115.2020v41i4a4
important” (KIM; BILLINGTON, 2016, p. 139). For Tiewtrakul and Fletcher (2010), “accent is one critical factor, especially where two non-native English speaking parties have to communicate using the English language” (TIEWTRAKUL; FLETCHER, 2010, p. 238). However, accent does not seem to be a critical factor just for NNES. Research has shown that even NES process information more slowly when listening to unfamiliar native accents under adverse listening conditions (ADANK et al., 2009).

Another factor that might interfere with the ease of understanding an ATCO’s transmission is the density of the transmission. Research has shown that information density has an impact on communication performance (ESTIVAL; MOLESWORTH, 2016; TIEWTRAKUL; FLETCHER, 2010). Although it has been recommended that the pilot/air traffic control transmissions should not include more than three pieces of information (BARSHI, 1997, as cited in WU et al, 2019), transmissions sometimes include much more than that. For example, Wu et al. (2019) counted up to eight items in a single transmission. Estival and Molesworth’s (2016) study indicated that “the increased number of items per transmission makes it more difficult for pilots to read back accurately what they have been told” (ESTIVAL; MOLESWORTH, 2016, p. 225), especially for NNES pilots and less experienced NES pilots. Further, issues regarding sound quality and distortion are common features of communication between pilots and ATCOs. Early work by Spolsky et al. (1968) found correlations between proficiency and the ability to understand distorted messages – the so-called ‘white noise’ tests.

Finally, aviation professionals may come across a wide variety of unexpected or unfamiliar topics, and this might also have a negative impact on their ability to understand what they hear. Research has shown that topic familiarity impacts language listening comprehension, regardless of the listener’s proficiency level (SCHMIDT-RINEHART, 1994). This is another important factor that needs to be taken into consideration when identifying the listening comprehension construct for test design and task specification in the aeronautical English workplace.

2.3 The assessment of listening in the context of pilot/ATCOs’ RT communications

The ICAO LPRs call for the assessment of plain English. Garcia (2015) suggested that tests should also include the assessment of phraseology. As one of the participants in her study explained:

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 … 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. (GARCIA, 2015, p. 34)

Previous research has also shown that the ICAO rating scale descriptors for comprehension are deficient. Knoch (2009) argued that “comprehension could not be accurately measured in a scale designed to assess speaking performance” (KNOCH, 2009, p. 31). Similarly, Kim and Billington (2016) claimed that the comprehension descriptors in the ICAO rating scale seemed to be inadequate, as the ability to comprehend a range of accents is not mentioned in Level 4. Also, in Pfeiffer’s (2009) study, the lowest inter-rater reliability result among the six criteria was the one for the assessment of comprehension. She suggested the reason for this might have been that it is difficult to assess some of the descriptors for comprehension. According to Pfeiffer (2009), “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” (PFEIFFER, 2009, p. 56). She adds that the comprehension descriptors were “not very enlightening” and that, in her judgment, “they are possibly the least well thought out in the entire rating scale” (PFEIFFER, 2009, p. 57).

As evidenced in the discussion above of literature relevant to the language assessment of pilots’ and ATCOs’ listening comprehension in the aeronautical English workplace, there is still a lack of clarity with regard to assessment needs. Thus, this study is guided by the following research questions:

1) Based on accounts of key stakeholders, what needs to be assessed in relation to the listening performed by pilots in RT communications with the ATCOs?

2) Based on accounts of key stakeholders, what are some of the implications of these assessment needs for the development of listening tasks?

Any definitive answers to these questions are exceptionally complex and beyond the scope of the present study. Our intention has been to conduct a preliminary needs analysis of the language that should be targeted in order to help test developers to design tests and test tasks which may better represent the TLU domain (BACHMAN; PALMER, 2010; WAGNER, 2014). Douglas (2010) refers to needs analysis as “the process of deciding what and how to test” (DOUGLAS, 2010, p. 39). Although the present study focuses on “what to test”, “how to test” is considered in a larger study, of which the current study is a part.

3. Methods

http://revistas.pucsp.br/esp

DOI: https://doi.org/10.23925/2318-7115.2020v41n4a4
3.1 Overall research design

This mixed-methods study applied a two-phase explanatory sequential research design (CRESWELL, 2015). Having both quantitative and qualitative data, as Creswell (2014) argued, “provides a stronger understanding of the research problem or questions” (CRESWELL, 2014, p. 264). In the first phase of the study, quantitative data (QUAN) were drawn, along with some qualitative data (qual), through an online questionnaire using Qualtrics, a survey software. In the second phase, in-depth qualitative data were collected through semi-structured interviews. The qualitative data were gathered in order to help understand the quantitative results and elaborate on them. Ethics approval was granted by the Carleton University Research Ethics Board.

3.2 Participants

Jacoby and McNamara (1999), Kim and Elder (2009), and Knoch (2009) argued that LSP test developers, when developing criteria for rating, or when conducting validation research, should consider indigenous assessment (defining the assessment criteria based on domain experts’ understanding of communicative effectiveness) because, as Jacoby and McNamara argued, “language and communication may mean one thing to linguistically oriented professionals and another to gatekeepers within a professional community” (JACOBY; MCNAMARA, 1999, p. 236). Douglas (2001) also argues that an analysis of the specific language use situation, together with indigenous assessment criteria, should be the base of LSP assessment criteria. Fox and Artemeva (2017) also emphasized the importance of incorporating indigenous criteria in the development of tests, as, according to them, it likely makes tasks more meaningful and useful.

Taking this into consideration, participants in Phase 1 of the present study were airplane and helicopter pilots. 156 pilots participated in the study (151 men and 5 women). All participants were NNES. Seven were private pilots, 66 commercial pilots, 82 airline transport pilots and one was a military pilot. Among them, 145 participants were airplane pilots, four helicopter pilots and seven were both airplane and helicopter pilots. The majority of participants’ first language was Brazilian Portuguese (124), but other first languages included Spanish (6), Azerbaijan (1), French (1), and Thai (1). The participants’ experience with international flights (or outside their own country) can be seen in Table 1.

| Table 1. Participants’ experience with international flights |

http://revistas.pucsp.br/esp

DOI: https://doi.org/10.23925/2318-7115.2020v41i4a4
<table>
<thead>
<tr>
<th>Experience with international flights</th>
<th>Number of participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Never flown internationally</td>
<td>55</td>
</tr>
<tr>
<td>Up to 50 hours of international flights</td>
<td>17</td>
</tr>
<tr>
<td>51-200 hours of international flights</td>
<td>20</td>
</tr>
<tr>
<td>201-500 hours of international flights</td>
<td>12</td>
</tr>
<tr>
<td>501-1000 hours of international flights</td>
<td>8</td>
</tr>
<tr>
<td>1001-5000 hours of international flights</td>
<td>30</td>
</tr>
<tr>
<td>Over 5001 hours of international flights</td>
<td>14</td>
</tr>
</tbody>
</table>

139 participants reported having had their aeronautical English formally assessed according to the ICAO’s LPRs. Table 2 shows the final levels they were awarded.

**Table 2.** Reported final levels awarded to participants on aeronautical English language proficiency tests

<table>
<thead>
<tr>
<th>Final Level</th>
<th>Number of participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 – Pre-operational</td>
<td>10</td>
</tr>
<tr>
<td>4 – Operational</td>
<td>72</td>
</tr>
<tr>
<td>5 – Extended</td>
<td>42</td>
</tr>
<tr>
<td>6 – Expert</td>
<td>15</td>
</tr>
<tr>
<td>Not assessed</td>
<td>17</td>
</tr>
<tr>
<td>Total</td>
<td>156</td>
</tr>
</tbody>
</table>

In Phase 2 of the study there were six participants, four men and two women. All were NNES: four participants came from Brazil, one from France (but had been living in an Anglosphere country for many years), and one from Azerbaijan. They had extensive professional background in language assessment and aeronautical English. At the time of the study,

- four pilots had experience flying internationally and had also worked as ICAO language proficiency test raters. One of them was not only a pilot and a rater, but also held a PhD in Linguistics and was an active researcher in the field;

- one air traffic controller had been working as an ICAO test rater for more than 10 years;

- one expert in aeronautical English had worked for 15 years as a test developer, rater trainer and test administrator, held a PhD in Applied Linguistics and was also an active researcher in the field.

Participants were recruited through email and social media to participate in both phases of the study. Additionally, pilots who in the questionnaire reported having experience as ICAO test raters, were invited to participate in Phase 2.

### 3.3 Type of data and data collection methods

Alderson et al. (1995) suggested that a needs analysis might involve “questionnaires or interviews with language users, or the consultation of relevant literature or of experts on the type of communication involved” (ALDERSON et al., 1995, p. 22). In Phase 1, the quantitative data were gathered through an [http://revistas.pucsp.br/esp](http://revistas.pucsp.br/esp) DOI: [https://doi.org/10.23925/2318-7115.2020v4i4a4](https://doi.org/10.23925/2318-7115.2020v4i4a4)
online anonymous questionnaire. The questionnaire comprised of 19 questions (nine demographic questions, five closed-ended opinion questions, and one open-ended question at the end, which invited participants to offer comments of their own regarding the topics discussed in the questionnaire). Participants took an average of approximately 18 minutes to complete the questionnaire (Attachment A). It is important to mention that, although participants were required to finish the questionnaire in order to have the data saved at the end, they were not forced to answer all questions. For this reason, for some questions there were fewer than the total of 156 responses.

In Phase 2, one-on-one semi-structured voice-only interviews were carried out through Skype to gather rich qualitative data. All the interviews were conducted in English, and included some demographic questions, a general question on listening assessment (“what do you think should be assessed in terms of pilots’ listening comprehension in the radiotelephony communications?”), followed by the same questions that were asked in the questionnaire. Participants were requested to explain their answers, by answering questions such as “why?”, “why not?”, or “how?”. The duration of the interviews varied from approximately 40 to 75 minutes, with an average of approximately 53 minutes. The interviews were audio-recorded. Participants were assigned alphabetical codes (A to F) for confidentiality. The recordings were transcribed for analysis.

3.4 Methods of data analysis

Descriptive statistical analyses were carried out to investigate the quantitative data. The results of Phase 1 were used to inform Phase 2. To analyze the qualitative data, descriptive coding was performed (Saldaña, 2013), which “summarizes in a word or short phrase – most often as a noun – the basic topic of a passage of qualitative data” (SALDAÑA, 2013, p. 70). The codes that reflected the topics of the questions were: speech rate, accents, unfamiliar words, cultural subtleties, information density, phraseology, complex grammar, main ideas, recognition of words and numbers, and extract specific meaning. Some other generated codes represent topics that were not explicitly discussed, but that were recurrent during the interviews. They were: background noise, radio interference, frequency congestion, workload, expectations, and interactions. In order to validate the coding, both authors performed the coding and consensus was reached. The findings from the quantitative and qualitative phases were integrated and are reported in the next section.

4. Results and Discussion

http://revistas.pucsp.br/esp

DOI: https://doi.org/10.23925/2318-7115.2020v41i4a4
From the analysis of both qualitative and quantitative data, it is possible to identify what stakeholders perceive as key listening comprehension needs that should be included in the construct definition of a proficiency test that aims to assess pilots’ listening in the context of RT communications. Findings are summarized below.

4.1 The need to understand speech at different rates of delivery

79.49% of participants in the survey (n=124) stated that a faster ATCO speech rate interferes with the ease of understanding an ATCO’s transmission and that assessment based on varying speeds of delivery is either important (n=90) or extremely important (n=43), with a mean response of 5.05. Table 3 summarizes the pilots’ responses to the importance of understanding speech at different rates of speed.

<table>
<thead>
<tr>
<th>Ability</th>
<th>Extremely irrelevant (1)</th>
<th>Irrelevant (2)</th>
<th>A little irrelevant (3)</th>
<th>A little important (4)</th>
<th>Important (5)</th>
<th>Extremely important (6)</th>
<th>Total</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>To Understand speech at different speed</td>
<td>0 (0.00%)</td>
<td>4 (2.56%)</td>
<td>4 (2.56%)</td>
<td>15 (9.62%)</td>
<td>90 (57.69%)</td>
<td>43 (27.56%)</td>
<td>156</td>
<td>5.05</td>
</tr>
</tbody>
</table>

In the survey, participants were asked to listen to four extracts of the same air traffic control message spoken at four different speeds (100, 140, 180, and 220 words per minute [wpm]) and to answer how frequently they hear ATCOs speaking at each of the given speeds. Table 4 shows their answers. By looking at the means, we can see that the speech rate participants reported hearing the most was 180 wpm, and the one they claimed to hear the least was 100 wpm, which is the speech rate recommended by ICAO to be used. One of the survey participants commented that “controllers usually speak too fast”. These results confirm Prinzo and Britton’s (1993) findings that ATCOs speak very fast.

<table>
<thead>
<tr>
<th>Speech rate</th>
<th>Never (1)</th>
<th>Rarely (2)</th>
<th>Sometimes (3)</th>
<th>Frequently (4)</th>
<th>Always (5)</th>
<th>Total</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 wpm</td>
<td>61 (39.61%)</td>
<td>63 (40.91%)</td>
<td>13 (8.44%)</td>
<td>13 (8.44%)</td>
<td>4 (2.60%)</td>
<td>154</td>
<td>1.94</td>
</tr>
<tr>
<td>140 wpm</td>
<td>18 (11.69%)</td>
<td>40 (25.97%)</td>
<td>46 (29.87%)</td>
<td>43 (27.92%)</td>
<td>7 (4.55%)</td>
<td>154</td>
<td>2.88</td>
</tr>
<tr>
<td>180 wpm</td>
<td>4 (2.61%)</td>
<td>10 (6.54%)</td>
<td>27 (17.65%)</td>
<td>101 (66.01%)</td>
<td>11 (7.19%)</td>
<td>153</td>
<td>3.69</td>
</tr>
<tr>
<td>220 wpm</td>
<td>19 (12.42%)</td>
<td>41 (26.80%)</td>
<td>35 (22.88%)</td>
<td>43 (28.10%)</td>
<td>15 (9.80%)</td>
<td>153</td>
<td>2.96</td>
</tr>
</tbody>
</table>

Five interview participants agreed that ATCOs tend to speak faster than the recommended 100 wpm. For participant A, 100 wpm is not realistic or natural. The other participant said that only American

http://revistas.pucsp.br/esp

DOI: https://doi.org/10.23925/2318-7115.2020v41i4a4
ATCOs speak faster. For him, “if you call the control faster and with very good English the controller answers very quickly, but if you call the control in a slow way, the controller answers in the same way”. Participants B, C and F stated that ATCOs speak very fast especially in busier airspaces, where their workload is heavy. This supports Prinzo and Britton’s (1993) finding that people talk faster when under stress or increased mental workload. ATCOs should be trained to speak at the recommended speech rate. Participant B stated that the speed of delivery might be one of the main challenges that pilots face when trying to understand an ATCO’s transmission. According to her, “this is very important because, even if you have an accent in your pronunciation, but your speed is okay, you sound more intelligible. But when it is the other way around, it is very hard to understand.” The results corroborate Buck’s (2001) claim that when the speech rate is greater, understanding is more difficult.

4.2 The need to understand different accents

64 participants in the survey (43.54%) responded that they have spent more than 60% of their time while conducting international flights listening to NNES, 47 (31.97%) answered that they have spent between 40% and 60% of their time listening to NNES, and 36 (24.49%) reported to have spent less than 40% of their time listening to NNES. Although interviewees claimed that how much a pilot listens to a NES ATCO or to a NNES ATCO will depend on factors such as where they usually fly or the company they fly for, most of them reported that pilots usually speak more to NNES than to NES. As participant A pointed out, because of the numbers of NES and NNES around the world, “it makes sense that most of the interactions would be with non-native English speakers when they are on international flights”. For participant B, “it is common sense that we have more second language speakers of English communicating with each other than with first language English speakers”. On the other hand, participant C stated “there is a tendency that the pilots have more chance to speak with native speakers … because the traffic is denser in countries that speak English”. According to participant E, pilots spend 50% of their time talking to NNES and 50% to NES.

The survey participants claimed that they hear unintelligible pronunciation with the frequency described in Table 5. The mean was 3.13, which indicates that the average was between sometimes and frequently. As Buck (2001) pointed out, “when listeners hear an unfamiliar accent … this can cause problems and may disrupt the whole comprehension process. An unfamiliar accent can make comprehension almost impossible for the listener” (p.35). It is important to note that an unintelligible
pronunciation does not necessarily mean that it is a non-native speaker’s pronunciation. As a matter of fact, interview participants A and F reported that understanding a NNES accent might be easier than a NES one.

**Table 5.** How frequently do pilots hear unintelligible pronunciation in RT communications?

<table>
<thead>
<tr>
<th>How frequently pilots hear unintelligible pronunciation</th>
<th>Never (1)</th>
<th>Rarely (2)</th>
<th>Sometimes (3)</th>
<th>Frequently (4)</th>
<th>Always (5)</th>
<th>Total</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2 (1.28%)</td>
<td>30 (19.23%)</td>
<td>75 (48.08%)</td>
<td>44 (28.21%)</td>
<td>5 (3.21%)</td>
<td>156</td>
<td>3.13</td>
</tr>
</tbody>
</table>

72 participants in the survey (46.15%) reported that a difficult accent interferes with the ease of understanding an ATCO’s transmission. For five interview participants, it is common for pilots to struggle to understand ATCOs’ pronunciation, especially, as participant A said, when they are not familiar with the accent. Fortunately, as participant E pointed out, pilots eventually get used to the accents they hear. On the other hand, for participant D, pilots do not struggle so often because “if someone has a strong accent, in the context, it is easy to understand”.

Table 6 shows how much the pilots who participated in the survey considered it important to assess the ability to understand different accents. It is possible to see, by looking at the mean, that the average of the participants considered the assessment of the ability to understand different accents to be of some importance.

**Table 6.** How important is it to assess the ability to understand different accents?

<table>
<thead>
<tr>
<th>Ability</th>
<th>Extremely irrelevant (1)</th>
<th>Irrelevant (2)</th>
<th>A little irrelevant (3)</th>
<th>A little important (4)</th>
<th>Important (5)</th>
<th>Extremely important (6)</th>
<th>Total</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>To understand different accents</td>
<td>0 (0.00%)</td>
<td>7 (4.49%)</td>
<td>5 (3.21%)</td>
<td>31 (19.87%)</td>
<td>86 (55.13%)</td>
<td>27 (17.31%)</td>
<td>156</td>
<td>4.78</td>
</tr>
</tbody>
</table>

All six interview participants stated that it is important to assess pilots’ ability to understand different accents. Five of them argued that it is important for all pilots who are authorized to fly internationally (pilots who are Level 4 and above) to demonstrate this ability because they need to be prepared as they are going to listen to different accents. This corroborates Kim and Billington’s (2016) claim that “the criteria should explicitly require pilots and controllers to demonstrate comprehension ability across a broad range of accents, and that this should be included from Level 4” (KIM; BILLINGTON, 2016, p. 154).
4.3 The need to understand the meaning of uncommon words and expressions

51 participants in the survey (32.69%) reported that unknown vocabulary interferes with the ease of understanding an ATCO’s transmission. Table 7 shows that the average frequency of hearing words and expressions whose meaning pilots do not understand was between rare and occasional (sometimes).

**Table 7.** How frequently, in RT communications, do pilots hear words and expressions whose meaning they do not understand?

<table>
<thead>
<tr>
<th>How frequently pilots hear words and expressions whose meaning they do not understand</th>
<th>Never (1)</th>
<th>Rarely (2)</th>
<th>Sometimes (3)</th>
<th>Frequently (4)</th>
<th>Always (5)</th>
<th>Total</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7 (4.49%)</td>
<td>60 (38.46%)</td>
<td>73 (46.79%)</td>
<td>15 (9.62%)</td>
<td>1 (0.64%)</td>
<td>156</td>
<td>2.63</td>
</tr>
</tbody>
</table>

Table 8 shows how important the pilots who participated in the survey considered the assessment of the ability to understand the meaning of uncommon words and expressions. The mean of 4.34 indicates that on average participants in the study considered the assessment of the ability to understand the meaning of uncommon words and expressions to be of some importance.

**Table 8.** How important is it to assess the ability to understand the meaning of uncommon words and expressions?

<table>
<thead>
<tr>
<th>Ability</th>
<th>Extremely irrelevant (1)</th>
<th>Irrelevant (2)</th>
<th>A little irrelevant (3)</th>
<th>A little important (4)</th>
<th>Important (5)</th>
<th>Extremely important (6)</th>
<th>Total</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>To understand the meaning of uncommon words and expressions</td>
<td>1 (0.64%)</td>
<td>13 (8.33%)</td>
<td>17 (10.90%)</td>
<td>47 (30.13%)</td>
<td>57 (36.54%)</td>
<td>21 (13.46%)</td>
<td>156</td>
<td>4.34</td>
</tr>
</tbody>
</table>

Interview participants B, E and F considered it important to assess the understanding of unfamiliar vocabulary, whereas participants C and D did not. Participant A said that it is important only if the words and expressions are related to aviation, remarking, “there are things that we are not able to understand and that are not crucial”. Participant B said that it is important because “unusual situations may happen inside the plane or on the runways”. Participant F agreed that it is important because “even though everything is standardized, in the real life, real world, sometimes things don't happen as the manual says”. He further explained that “some airports have some restrictions that the pilots might not be familiar with and may not understand what they are not supposed to do”. One survey respondent also argued that “although the international standard phraseology suits the needs of air traffic on a normal day, even to the extent that multinational controller-pilot communication is facilitated by it, basic English vocabulary and plain English conversation capability prove to be an obstacle when a given situation goes abnormal”.

http://revistas.pucsp.br/esp

DOI: https://doi.org/10.23925/2318-7115.2020v41i4a4
Participants C and D, on the other hand, noted that, although pilots listen to words that they do not understand, it is not so important to assess this skill. For participant D, “the most difficult is to understand the name of the intersections. I can give you an example, we have an intersection here in Brazil called Utbur. And when American pilots says Utbur, it is completely different. And there are a lot of crazy names.” Although unfamiliar vocabulary is heard in radiotelephony, participants might have not considered it to be extremely important to be assessed because, as three interview participants pointed out, when a pilot listens to an unfamiliar word, that is not a major problem because they can ask for clarification. Participant D argued that “if you are not sure about an unfamiliar word, it is much better to ask for clarification because sometimes you may think one thing, but it is totally wrong”.

In relation to idioms and phrasal verbs, participant A explained that ATCOs use them in radiotelephony, although they should not. As a matter of fact, the greatest difficulty that one of the survey participants reported facing was “when the controller uses some local expressions.” Indeed, research has shown that ATCOs use idiomatic vocabulary when communicating with pilots (PRINZO et al., 2009). Prinzo et al. found that the use of slang has caused readback errors. In addition, Prinzo’s (2009) results suggested that “colloquialisms, slang; pleasantries, although well meaning, may cause problems for those who are less proficient in the English language” (PRINZO, 2009, p. 11). All interview participants reported that ATCOs use idiomatic vocabulary (three participants emphasized that it happens especially in the USA). As participant B pointed out

We have like a paradox in radiotelephony communications, because if everybody complied with the rules, the procedures and the standard phraseology, we wouldn't have pilots or controllers resorting to slang, to idioms or to more colloquial language over the radio, but sometimes this happens and I think more often than we would desire. … So, they need to be able to understand but not to use them in radiotelephony (Participant B).

As idiomatic vocabulary is part of the real world, two participants said that the understanding of phrasal verbs and idioms should be assessed. However, they both think that it should only be assessed in tests that assess Level 6 pilots.

4.4 The need to recognize indirect meaning

Table 9 shows how pilots who participated in the survey viewed the importance of assessing the ability to recognize indirect meaning (e.g., politeness, sarcasm, professionalism, anger). 51 participants considered this ability irrelevant, whereas 105 considered it important (the mean of 3.9 suggests most participants thought it was not of great importance).
Table 9. How important is it to assess the ability to recognize indirect meaning?

<table>
<thead>
<tr>
<th>Ability</th>
<th>Extremely irrelevant (1)</th>
<th>Irrelevant (2)</th>
<th>A little irrelevant (3)</th>
<th>A little important (4)</th>
<th>Important (5)</th>
<th>Extremely important (6)</th>
<th>Total</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>To recognize indirect meaning (e.g., politeness, sarcasm professionalism, anger)</td>
<td>3 (1.92%)</td>
<td>24 (15.38%)</td>
<td>24 (15.38%)</td>
<td>53 (33.97%)</td>
<td>38 (24.36%)</td>
<td>14 (8.97%)</td>
<td>156</td>
<td>3.90</td>
</tr>
</tbody>
</table>

Four interview participants remarked that it is important to assess pilots’ ability to understand cultural subtleties. As participant B argued, “we do not communicate without resorting to our cultural frames of reference”. Participants E and F indicated it should be tested in Level 6 tests. Participant B argued that even Level 4 speakers should demonstrate this ability. Participant A was not sure if it was important, but she stated that it should be included at least in the pilots’ training. Half of the interview participants agreed that the ability to recognize indirect meaning was important to assess, and half agreed it was not. Participant A asserted that it should be assessed only at Level 6. Participant B said interpreting indirect meaning was important because “they need to be able to recognize that in order to deal with the situations in a very professional way”. Participants who stated that it was not important argued that it is unusual for ATCOs to be sarcastic or angry, and that they are usually polite and professional. Although instances of unprofessional tone and attitude, such as impatience, sarcasm, impoliteness, and arrogance might not occur very frequently in RT communications (MONTEIRO, 2019), pilots and ATCOs who participated in Monteiro’s survey considered such instances to be potential threats to safety.

4.5 The need to understand transmissions with more than three items

For 81 of the participants in the survey (51.92%), a high number of pieces of information in a transmission interferes with the ease of understanding a transmission. Most participants reported that according to their experience, a single air traffic control (ATC) transmission might include more than the ICAO recommended three pieces of information per transmission. Table 10 shows the answers that the survey participants gave to the question regarding the maximum number of elements they have heard in a single ATC’s transmission. The mean was 2.74, which suggests that the average of the participants reported hearing a maximum of four to five items in an ATC transmission. One survey participant mentioned that “usually all controllers speak up to three pieces of information”. Another survey participant explained that “depending on the phase of the flight, there might be a higher number of pieces
of information in a message. When requesting a clearance before starting the engines, be prepared to write down a lot of information.” A different participant commented that “as a pilot, I can assume that it is difficult to listen to a high number of items of information during a critical phase of flight. The most critical phases are during the initial climb-out and approach to land.”

**Table 10.** What was the maximum number of pieces of information participants listened to in a single transmission from an ATCO?

<table>
<thead>
<tr>
<th>Maximum number of pieces of information in each communication</th>
<th>Number of participants (percentage)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>29 (19.08%)</td>
</tr>
<tr>
<td>4</td>
<td>43 (28.29%)</td>
</tr>
<tr>
<td>5</td>
<td>41 (26.97%)</td>
</tr>
<tr>
<td>6</td>
<td>24 (15.79%)</td>
</tr>
<tr>
<td>7</td>
<td>7 (4.61%)</td>
</tr>
<tr>
<td>8 or more</td>
<td>8 (5.26%)</td>
</tr>
</tbody>
</table>

All five interview participants who were pilots reported having heard more than three pieces of information in a single ATCO’s transmission. Participant A argued there might be up to eight items. Participant D claimed there might be up to 20 elements in a single transmission, but there are typically around three or four. He explained:

> During the clearance, taxi and reclearance, there are a lot of pieces of information. … In big airports there are a lot of taxiways, a lot of runways. […] It is very difficult to read back or take note because of the name of the airways, VORs, etc., so the most difficult situation is during those phases because there are a lot of different names, different numbers, different taxiways, and there is a lot of information (Participant D).

Participant C said there are usually around five, participant E said around three or four, sometimes more than that, and participant F reported around three to five.

Although ICAO recommends a maximum of three items per ATC transmission, the reality seems to be that pilots listen to ATCOs’ transmissions with more than three items. It is important to note that, although a test construct would need to represent the features of the real-life language use (and this also applies to the use of idiomatic vocabulary and the need to adhere to phraseology), ATCOs should be encouraged to comply with the recommendations, as Wu et al. (2019) also suggested. As participant C explained, “if you give a very long instruction, […] you could have a misunderstanding”. According to Wu et al. (2019), “the number of items in a message, known as information density, has also been found to increase error rate in pilot communications” (WU et al., 2019, p. 3). Additionally, participant F expressed concern in relation to the importance of human memory in this context. Buck (2001) and Wu (1998) also highlighted the role that memory plays in the listening process. According to the interviewee,

DOI: https://doi.org/10.23925/2318-7115.2020v4i4a4
“because sometimes pilots don’t have time to write down the information […] , they rely a lot on their memories and if you give them too much information, it is guaranteed that something is going to be missed”. One of the survey participants also commented on this issue: “it is easy to understand. The problem is to remember all the information. Even in Portuguese.” As Buck (2001) pointed out, a person whose processing is more controlled than automatic is a less efficient listener (Buck, 2001).

4.6 The need to understand both plain English and phraseology

The findings of the present study correspond with Garcia’s (2015), as both studies suggest that not only plain English but also phraseology should be assessed. The vast majority of the survey participants considered that it is very important to assess comprehension of both plain English and phraseology. Table 11 shows that 96.76% of participants somewhat agree, agree or strongly agree that it is important to assess comprehension of plain English, and 96.16% somewhat agree, agree or strongly agree that it is important to assess understanding of phraseology.

Table 11. How important is it to assess comprehension of plain English and phraseology?

<table>
<thead>
<tr>
<th></th>
<th>It is important to assess comprehension of plain English</th>
<th>It is important to assess understanding of phraseology</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of participants</td>
<td>Percentage</td>
</tr>
<tr>
<td>I strongly disagree with it</td>
<td>1</td>
<td>0.65%</td>
</tr>
<tr>
<td>I disagree with it</td>
<td>4</td>
<td>2.60%</td>
</tr>
<tr>
<td>I somewhat disagree with it</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>I somewhat agree with it</td>
<td>17</td>
<td>11.04%</td>
</tr>
<tr>
<td>I agree with it</td>
<td>65</td>
<td>42.21%</td>
</tr>
<tr>
<td>I strongly agree with it</td>
<td>67</td>
<td>43.51%</td>
</tr>
<tr>
<td>Total</td>
<td>156</td>
<td></td>
</tr>
</tbody>
</table>

All six interview participants agreed that it is important to assess comprehension of phraseology. Participant D remarked that assessing comprehension of phraseology is of utmost importance. As interview participant A pointed out, communication in aviation is based on “the comprehension of phraseology and the ability to produce phraseology correctly”, so it should be assessed. Participant B claimed that “what pilots face in their daily routines while they are communicating with air traffic controllers includes both phraseology and plain English. So, if this is part of their reality, […] it should be assessed because it is part of the construct.”.

http://revistas.pucsp.br/esp DOI: https://doi.org/10.23925/2318-7115.2020v41i4a4
4.7 The need to understand complex grammar

Only 26 participants in the survey (16.67%) responded that complex grammar interferes with the ease of understanding an ATCO’s transmission. Table 12 shows the responses given by the survey participants to the question of how frequently pilots hear sentences they cannot understand because of structure. The average respondent said they have this problem between rarely and sometimes.

**Table 12.** How frequently, in RT communications, do pilots hear sentences they cannot understand because of the way the words are put together (structure)?

<table>
<thead>
<tr>
<th>How frequently pilots hear sentences they cannot understand because of the way the words are put together (structure)</th>
<th>Never (1)</th>
<th>Rarely (2)</th>
<th>Sometimes (3)</th>
<th>Frequently (4)</th>
<th>Always (5)</th>
<th>Total</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12 (7.69%)</td>
<td>68 (43.59%)</td>
<td>60 (38.46%)</td>
<td>13 (8.33%)</td>
<td>3 (1.92%)</td>
<td>156</td>
<td>2.53</td>
</tr>
</tbody>
</table>

For four interview participants, pilots do not need to understand complex grammar. One mentioned that it is not important at least for level 4, and another one said that it is not important for Levels 4 and 5, but that it is for Level 6. Participant D argued that messages are very short and they can be summarized by cutting out words. Participant E claimed that pilots just need to understand the basic verb tenses like present simple, present continuous, past and future, but not future continuous or past perfect. Participant C, on the other hand, argued it is important. As he explained, “when the airport closes and the air traffic controller has to explain the reason why the airport is closed, a special situation, and to give some alternatives for the pilot to decide, I think the pilot should be required to understand complex grammar”, and participant B claimed that, although ATCOs are not supposed to use complex grammar, they sometimes do, so it might be important for pilots to understand it.

4.8 Other important needs

Results suggest other important needs. Table 13 shows how the pilots who participated in the survey viewed the importance of assessing the abilities to understand main ideas, to recognize words and numbers, and to extract specific meaning from both short and long transmissions. By looking at the means, we can see that the average of the participants in the survey considered the assessment of the ability to recognize words and numbers, to understand main ideas, and to extract specific meaning from both short and long transmissions, to be between important and extremely important.
Table 13. How important is it to assess the ability to understand main ideas, the ability to recognize words and numbers and the ability to extract specific meaning from both short and long transmissions?

<table>
<thead>
<tr>
<th>Abilities</th>
<th>Extremely irrelevant (1)</th>
<th>Irrelevant (2)</th>
<th>A little irrelevant (3)</th>
<th>A little important (4)</th>
<th>Important (5)</th>
<th>Extremely important (6)</th>
<th>Total</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>To understand main ideas</td>
<td>1 (0.64%)</td>
<td>0 (0.00%)</td>
<td>1 (0.64%)</td>
<td>1 (0.64%)</td>
<td>59 (37.82%)</td>
<td>94 (60.26%)</td>
<td>156</td>
<td>5.56</td>
</tr>
<tr>
<td>To recognize words and numbers</td>
<td>0 (0.00%)</td>
<td>0 (0.00%)</td>
<td>0 (0.00%)</td>
<td>2 (1.28%)</td>
<td>38 (24.36%)</td>
<td>116 (74.36%)</td>
<td>156</td>
<td>5.73</td>
</tr>
<tr>
<td>To extract specific meaning from both short and long transmissions</td>
<td>0 (0.00%)</td>
<td>2 (1.29%)</td>
<td>1 (0.65%)</td>
<td>14 (9.03%)</td>
<td>84 (54.19%)</td>
<td>54 (34.84%)</td>
<td>155</td>
<td>5.21</td>
</tr>
</tbody>
</table>

Of all the abilities included in the survey items, recognizing numbers was considered the most important by the pilots in this study, as the mean was 5.73. Interview participant B claimed that pilots “have to understand lots of information involving numbers, such as runway numbers, altitudes, speed, directions, and headings.” Participant F expressed concern in relation to the comprehension of numbers and what they refer to. As he explained:

The most important thing would be numbers and the ability to distinguish them from flight levels or headings or even air speed, as pilots usually get confused with headings and altitude. That is very dangerous. So, I think that the main thing would be precise numbers related to what they are meant to be, for example, the pilot is cleared for flight level 050, and then sometimes we can hear the candidate saying that they are supposed to fly on heading 050 (Participant F).

For participant D, it is very important to understand numbers and aviation words. Recognizing words also seems to be a very important skill. Participant C argued that understanding isolated words is crucial, especially the ones that are more critical to safety, but also it is very important to understand the main idea. Participant B explained that:

They need to recognize some isolated information like numbers, letters and some important words, but they also need to understand the meaning of certain transmissions which is broader. … For example, if the controller is informing about an emergency situation or unusual situation happening in an airport, or if they are in an emergency situation reporting a problem and the controller asks questions to get more details about this problem (Participant B).

For participant A, it is important for the pilot to be able to extract important information that is presented. As she stated, “some words, some bits of information are not that important, and you don’t need to retain that in your head, whereas others are absolutely crucial and you must be able to extract them”. She argued
that it is not only important for pilots to extract important information from what is said, but also to understand the cues given by prosodic features of pronunciation, such as intonation and stress.

For interview participant B, pilots’ ability to understand ATCOs “also depends on technical issues, like background noise or radio interference, poor radio quality. Sometimes they face difficulties in understanding. It is not just because they lack the abilities but sometimes this can be exacerbated by other factors.” For 104 participants in the survey (66.67%), bad quality of the radio transmission interferes with the ease of understanding. This supports Prinzo’s (2009) findings that indicated that technical issues such as radio interference, noise and problems in the transmission contribute to communication issues. Participant B pointed out that the quality of the radio transmission may be a big challenge for pilots. However, low quality of the transmission seems to be less of a problem nowadays than it used to be in the past. Participant D stated that “sometimes the radio has a lot of interference”. However, he also said that “the quality of the radios I think nowadays is very clear. I don't have problems with this.” Participant E also argues that the quality of the transmission is “much better than it used to be” and that “you get used to this distortion”. In relation to the background noise, as Estival and Molesworth (2009) explained, “the cockpit is a noisy environment and the presence of noise makes it more difficult for [NNES] listeners to understand speech” (ESTIVAL; MOLESWORTH, 2009, p. 24.4). Indeed, research has shown that the more noise, the harder it is to understand speech (SHIMITZU et al., 2002, as cited in ESTIVAL; MOLESWORTH, 2009).

Another technical issue that has been discussed in the interviews and that might interfere with the ease of understanding an ATCO’s transmission was radio frequency congestion. Although results from previous research have suggested that “radio frequency congestion does not affect pilots’ ability to communicate effectively, irrespective of their language background, pilot qualification or flying experience” (ESTIVAL; MOLESWORTH, 2016, p. 232), participants have argued that it may have an impact. For participant D, congestion can make communication very confusing as the transmission may become “very blocked”. Additionally, participant A pointed out that pilots need to have the ability to understand when the ATCO is addressing them. As she explained, “on the frequency, when there are a lot of other aircraft, and there is one communication that is directed to you, you want people to be able to recognize it, and respond. ... I think that is really important in testing comprehension. Being able to identify ... what is actually relevant to you and what is addressed to you.” Cushing (1994) gave real examples of this kind of miscommunication which he called an “uncertain addressee” problem of reference (when a pilot complied with a descent clearance intended for another aircraft).

http://revistas.pucsp.br/esp

DOI: https://doi.org/10.23925/2318-7115.2020v4i4a4
Another recurrent theme was that pilots need to have the ability to deal well with their own expectations. Two survey participants commented on this, as well as five interview participants. One survey participant said that most of what the ATCO tells the pilot before starting the engines, pilots already expect to hear, “but sometimes controllers change the requested route, flight level or even an expected departure procedure.” Another respondent said that most miscommunications happen “due to some out of script requests or information being passed”. Participant A commented that “they expect something and, if the air traffic controller doesn't tell them what they think or they are expecting, that is going to be an issue.” Participant F argued that “the main difficulties are when the pilot is expecting something, and, out of nowhere, something changes, and they get instructions that are totally different from what they were expecting to hear”. According to participant E, “you should be very vigilant” when expecting instructions from the ATCO in order not to have a miscommunication. Thus, pilots need to be aware of their expectations in order to be able to identify when the situation does not happen according to what they were expecting. It seems to be important to assess this ability.

A theme that has emerged from three interviews was the need for pilots to be able to understand communications while having an increased workload or when under stress. Participant C argued that “the pilot has lots of things to do inside the cockpit so if the air traffic control gives him much information, he won't be able to process this information”. For participants B and E, coping with heavy workload is a challenge for pilots.

A recurrent theme in the interviews was the need for pilots to interact well. All interview participants emphasized the importance of pilots being able to use strategies in order to ensure accurate communication. As Wagner (2014) pointed out, assessing listening together with speaking is challenging, “but it is a necessary and advisable goal” (WAGNER, 2014, p. 6). Participant A discussed the importance of testing “the ability to respond appropriately”, and participant E talked about the importance of assessing “how the pilots receive and respond”. As participant E argued, “if you do not understand something, it is much better to ask to repeat the information”. Participant C pointed out that “they won't be able to know everything, so they need to develop skills that can lead them to a different way to understand what is going on if they don't understand a word, so they have to be skilled at asking for repetition and paraphrasing”. For participant F, not understanding a transmission is not a major problem as long as the pilot is able to interact well. As he explained, “if a pilot is not capable of understanding, I am pretty sure that he is capable, at least, to say that, ‘hey, I'm not a great speaker of this language. So please calm down. I need

http://revistas.pucsp.br/esp

DOI: https://doi.org/10.23925/2318-7115.2020v41i4a4
just the basic information’’. Field (2019) argued that a competent listener must be “capable of using strategies effectively to compensate for gaps in understanding” (FIELD, 2019, p. 14). Both NES and NNES pilots and ATCOs need to be trained in order to learn strategies that will help them to interact appropriately and effectively. The importance of pilot training was emphasized by four interview participants. A survey participant complained that “native speakers usually don’t care if listeners are native or not”. As interview participant F pointed out, NNES “try to use simple words for conversations, and they do make an effort to understand you and to be understood. Sometimes native speakers just don't have that kind of feeling. They speak the way they are used to, and they don't care about how to express themselves, so they are not worried about it.” Participant A also claimed that “some air traffic controllers don't necessarily try to make it as clear as they should.” It is important to note that three interview participants mentioned the difficulty in understanding American ATCOs. They suggested that this might be the case because American ATCOs do not seem to follow the standardized phraseology. It seems to be important for pilots and ATCOs to have the ability to accommodate to less proficient users of the language. As argued by O’Malley et al. (1989), learners can be taught to apply some strategic resources which can help them to become more effective listeners.

**Conclusion**

This study responded to two research questions that explored key stakeholders’ perceptions of what needs to be assessed with regard to listening comprehension:

1) Based on accounts of key stakeholders, what needs to be assessed in relation to the listening performed by pilots in RT communications with the ATCOs?

2) Based on accounts of key stakeholders, what are some of the implications of these assessment needs for the development of listening tasks?

Investigating stakeholders’ needs may help to inform test development and support the definition of clear, coherent and well-articulated test constructs and concomitant specifications, thereby improving the overall technical quality of the required proficiency testing in this specialized domain. The second research question examined the implications of these assessment needs in developing assessment tasks. Table 14 provides a summary of the needs that were identified as an outcome of the present study and the implications for the development of listening assessment tasks.

http://revistas.pucsp.br/esp

DOI: https://doi.org/10.23925/2318-7115.2020v41i4a4
Table 14. Summary of pilots’ listening comprehension needs and implications for testing.

<table>
<thead>
<tr>
<th>Needs</th>
<th>Implications for the development of listening assessment tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>To understand speech at different speeds</td>
<td>The speech rate of the recordings should vary from 100 to 220 wpm. Most of them should be 180 wpm. At the same time, ATCOs should be trained to speak slower.</td>
</tr>
<tr>
<td>To understand different accents</td>
<td>Level 4 pilots should be tested on their ability to understand different accents, not only from Level 5 up. About half of the recordings should be made by NES and half by NNES. Only some accents (from both NES and NNES) should be more challenging.</td>
</tr>
<tr>
<td>To understand meaning of uncommon words and expressions</td>
<td>It should be tested, but it does not seem to be very important, as pilots have the benefit of asking for clarification when they do not understand a word. Idiomatic vocabulary should be assessed in tests designed to test Level 6.</td>
</tr>
<tr>
<td>To recognize indirect meaning</td>
<td>It seems to be important to be tested at least in tests designed to assess Level 6 pilots. It seems to be very important to be trained.</td>
</tr>
<tr>
<td>To understand transmissions with more than 3 items</td>
<td>Some recordings may have more pieces of information, but the average number should be around three or four.</td>
</tr>
<tr>
<td>To understand both plain English and phraseology</td>
<td>Understanding of not only plain English but also phraseology should be assessed.</td>
</tr>
<tr>
<td>To understand complex grammar</td>
<td>It does not seem to be very important to be tested. However, it should be tested in tests designed to assess Level 6 pilots.</td>
</tr>
<tr>
<td>To understand main ideas</td>
<td>It seems to be very important to test this ability.</td>
</tr>
<tr>
<td>To recognize words and numbers</td>
<td>It seems to be very important to test this ability.</td>
</tr>
<tr>
<td>To extract specific meaning from both short and long transmissions</td>
<td>It seems to be very important to test this ability.</td>
</tr>
<tr>
<td>To understand low quality transmissions</td>
<td>Low quality transmission does not seem to be a major problem anymore. Some interference may be included, but it should not be very serious.</td>
</tr>
<tr>
<td>To understand transmissions in spite of background noise</td>
<td>Aircraft noise should be included in the recordings as background noise.</td>
</tr>
<tr>
<td>To understand transmissions during increased workload</td>
<td>How to assess this ability should be further investigated.</td>
</tr>
<tr>
<td>To understand transmissions in congested radio frequency</td>
<td>How to assess this ability should be further investigated. It seems to be important for pilots to be able to recognize when they are being addressed.</td>
</tr>
<tr>
<td>To understand messages that differ from what was expected</td>
<td>It seems to be important to test this ability. Research should be conducted to find out how to do this.</td>
</tr>
<tr>
<td>To use strategies in order to understand</td>
<td>It seems to be very important to test this ability. It is also very important to be trained.</td>
</tr>
</tbody>
</table>

Some limitations of this study should be acknowledged. Firstly, the study only shows the perceptions of NNES who are mainly native Portuguese speakers, most of whom were residing in Brazil at the time of the study. Additionally, 35% of the survey participants reported that they had not flown internationally. Their opinions were based on their general knowledge and on their experience with domestic flights. More importantly, the generated data need to be treated as an indirect reflection of the participants’ understanding of the issues (what the participants claim may be different from the reality or even from what participants actually think).

http://revistas.pucsp.br/esp

DOI: https://doi.org/10.23925/2318-7115.2020v41i4a4
With these caveats in mind, it is useful to report some other themes that emerged from the qualitative data that warrant further exploration, and are currently under investigation in a large-scale study: the types of recordings that should be included in the test (authentic audios versus scripted studio recordings); the topics that the recordings should cover; how much interference and background noise to include in the recordings; how many times test takers should be allowed to listen to the recordings; how cultural subtleties can be assessed; the importance of assessing comprehension separately and also during real interactions; how increased workload can affect pilots’ ability to understand ATCOs’ transmissions; how to include in the listening test the challenges posed by traffic congestion; the need to revise the ICAO rating scale and what categories to possibly include in a scale to assess comprehension specifically. Further, additional research should also be conducted in order to identify the most appropriate listening tasks, test methods and procedures.

As Read and Knoch (2009) pointed out, “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” (READ; KNOCH, 2009, p. 21.10). Although there is still a lot to be investigated in relation to the assessment of the English used by pilots and ATCOs in the context of RT communications, it is hoped that this study will contribute to improving flight safety by helping language test developers to design tests and tasks that better represent the listening comprehension construct.

**Angela Carolina de Moraes Garcia** is in her first year of the PhD program in Applied Linguistics and Discourse Studies at Carleton University. She has been working as a language test developer and raters’ trainer at the Brazilian Civil Aviation National Agency (ANAC) since January 2008. Angela holds a Master’s degree in Language Testing from Lancaster University. Her research interests include English for Specific Purposes and Language Testing, especially aeronautical English testing, validity, reliability, construct definition and rating scale development. E-mail: angelagarcia@gmail.carleton.ca

**Janna Fox**, PhD, is Emeritus Professor of Applied Linguistics, in the School of Linguistics and Language Studies at Carleton University, Ottawa, Canada. Her current research interests include validation theory, language testing, classroom based assessment, and pre-service/in-service teacher development. Email: Janna.Fox@carleton.ca

**REFERENCES**

[http://revistas.pucsp.br/esp](http://revistas.pucsp.br/esp)  DOI: [https://doi.org/10.23925/2318-7115.2020v41i4a4](https://doi.org/10.23925/2318-7115.2020v41i4a4)


dehttp://dx.doi.org/10.1017/CBO9780511732959


GARCIA, A. C. M. 2015. 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. Master’s dissertation. Lancaster University, Lancaster, UK.


http://revistas.pucsp.br/esp

DOI: https://doi.org/10.23925/2318-7115.2020v41i4a4


http://revistas.pucsp.br/esp

DOI: https://doi.org/10.23925/2318-7115.2020v41i4a4
WU, Q.; B. MOLESWORTH; D. ESTIVAL. 2019. An investigation into the factors that affect miscommunication between pilots and air traffic controllers in commercial aviation. The International Journal of Aerospace Psychology, p. 53-63.

http://revistas.pucsp.br/esp DOI: https://doi.org/10.23925/2318-7115.2020v41i4a4
## Attachment A – Questionnaire

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
</table>
| 1 | I identify my gender as:  
( ) man ( ) woman ( ) non-binary  
Other: _ ______ ( ) do not wish to respond |
| 2 | Are you a native or a non-native speaker of English?  
I am a native speaker of English  
I am a non-native speaker of English |
| 3 | What is your first language?  
My first language is ________ |
| 4 | Have you ever had your aviation English formally assessed according to the ICAO (International Civil Aviation Organization) Language Proficiency Requirements?  
( ) yes ( ) no |
| 5 | (If ‘yes’ in question 4)  
What final Level were you awarded?  
1 – Pre-elementary  
2 - Elementary  
3 – Pre-operational  
4 - Operational  
5 - Extended  
6 - Expert |
| 6 | What aircraft do you fly?  
a) Airplane.  
b) Helicopter.  
c) Both airplane and helicopter. |
| 7 | What license do you hold?  
a) I am a private pilot.  
b) I am a commercial pilot.  
c) I am an airline transport pilot.  
d) I am a military pilot.  
e) I am not a pilot (If letter “f” was chosen, the survey would be ended) |
| 8 | How many hours have you flown internationally?  
a) I have never flown internationally  
b) Up to 50 hours of international flights  
c) 51-200 hours of international flights  
d) 201-500 hours of international flights  
e) 501-1000 hours of international flights  
f) 1001-5000 hours of international flights  
g) Over 5001 hours of international flights |
| 9 | To what extent do you agree or disagree with the following statement?  
“It is important to assess understanding of phraseology.”  
( ) I strongly disagree with it  
( ) I disagree with it  
( ) I somewhat disagree with it  
( ) I somewhat agree with it  
( ) I agree with it  
( ) I strongly agree with it |
<p>| 10 | Knowing that Plain English is the English used in radiotelephony communications when phraseology does not suffice, to what extent do you agree or disagree with the following |</p>
<table>
<thead>
<tr>
<th>Statement? “It is important to assess comprehension of plain English.”</th>
</tr>
</thead>
<tbody>
<tr>
<td>( ) I strongly disagree with it</td>
</tr>
<tr>
<td>( ) I disagree with it</td>
</tr>
<tr>
<td>( ) I somewhat disagree with it</td>
</tr>
<tr>
<td>( ) I somewhat agree with it</td>
</tr>
<tr>
<td>( ) I agree with it</td>
</tr>
<tr>
<td>( ) I strongly agree with it</td>
</tr>
</tbody>
</table>

11 To what extent do you consider it to be important to assess the pilots’ abilities in radiotelephony communications listed below?

<table>
<thead>
<tr>
<th>a</th>
<th>Understand main ideas</th>
</tr>
</thead>
<tbody>
<tr>
<td>( ) Extremely irrelevant</td>
<td></td>
</tr>
<tr>
<td>( ) Irrelevant</td>
<td></td>
</tr>
<tr>
<td>( ) A little irrelevant</td>
<td></td>
</tr>
<tr>
<td>( ) A little important</td>
<td></td>
</tr>
<tr>
<td>( ) Important</td>
<td></td>
</tr>
<tr>
<td>( ) Extremely important</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>b</th>
<th>Recognize words and numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>( ) Extremely irrelevant</td>
<td></td>
</tr>
<tr>
<td>( ) Irrelevant</td>
<td></td>
</tr>
<tr>
<td>( ) A little irrelevant</td>
<td></td>
</tr>
<tr>
<td>( ) A little important</td>
<td></td>
</tr>
<tr>
<td>( ) Important</td>
<td></td>
</tr>
<tr>
<td>( ) Extremely important</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>c</th>
<th>Understand different accents</th>
</tr>
</thead>
<tbody>
<tr>
<td>( ) Extremely irrelevant</td>
<td></td>
</tr>
<tr>
<td>( ) Irrelevant</td>
<td></td>
</tr>
<tr>
<td>( ) A little irrelevant</td>
<td></td>
</tr>
<tr>
<td>( ) A little important</td>
<td></td>
</tr>
<tr>
<td>( ) Important</td>
<td></td>
</tr>
<tr>
<td>( ) Extremely important</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>d</th>
<th>Recognize indirect meaning (e.g., politeness, sarcasm, professionalism, anger, etc.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( ) Extremely irrelevant</td>
<td></td>
</tr>
<tr>
<td>( ) Irrelevant</td>
<td></td>
</tr>
<tr>
<td>( ) A little irrelevant</td>
<td></td>
</tr>
<tr>
<td>( ) A little important</td>
<td></td>
</tr>
<tr>
<td>( ) Important</td>
<td></td>
</tr>
<tr>
<td>( ) Extremely important</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>e</th>
<th>Extract specific meaning from both short and long transmissions</th>
</tr>
</thead>
<tbody>
<tr>
<td>( ) Extremely irrelevant</td>
<td></td>
</tr>
<tr>
<td>( ) Irrelevant</td>
<td></td>
</tr>
<tr>
<td>( ) A little irrelevant</td>
<td></td>
</tr>
<tr>
<td>( ) A little important</td>
<td></td>
</tr>
<tr>
<td>( ) Important</td>
<td></td>
</tr>
<tr>
<td>( ) Extremely important</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>f</th>
<th>Understand speech at different speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>( ) Extremely irrelevant</td>
<td></td>
</tr>
<tr>
<td>( ) Irrelevant</td>
<td></td>
</tr>
<tr>
<td>( ) A little irrelevant</td>
<td></td>
</tr>
<tr>
<td>( ) A little important</td>
<td></td>
</tr>
<tr>
<td>( ) Important</td>
<td></td>
</tr>
<tr>
<td>( ) Extremely important</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>g</th>
<th>Understand the meaning of uncommon words and expressions</th>
</tr>
</thead>
<tbody>
<tr>
<td>( ) Extremely irrelevant</td>
<td></td>
</tr>
<tr>
<td>( ) Irrelevant</td>
<td></td>
</tr>
<tr>
<td>( ) A little irrelevant</td>
<td></td>
</tr>
<tr>
<td>( ) A little important</td>
<td></td>
</tr>
<tr>
<td>( ) Important</td>
<td></td>
</tr>
<tr>
<td>( ) Extremely important</td>
<td></td>
</tr>
</tbody>
</table>

12 How frequently, in radiotelephony, do you hear ____________?

<table>
<thead>
<tr>
<th>a</th>
<th>Unintelligible pronunciation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( ) Never</td>
<td></td>
</tr>
<tr>
<td>( ) Rarely</td>
<td></td>
</tr>
<tr>
<td>( ) Sometimes</td>
<td></td>
</tr>
<tr>
<td>( ) Frequently</td>
<td></td>
</tr>
<tr>
<td>( ) Always</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>b</th>
<th>Words and expressions whose meaning you do not understand</th>
</tr>
</thead>
<tbody>
<tr>
<td>( ) Never</td>
<td></td>
</tr>
<tr>
<td>( ) Rarely</td>
<td></td>
</tr>
<tr>
<td>( ) Sometimes</td>
<td></td>
</tr>
<tr>
<td>( ) Frequently</td>
<td></td>
</tr>
<tr>
<td>( ) Always</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>c</th>
<th>Sentences you cannot understand because of the way the words are put together (structure)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( ) Never</td>
<td></td>
</tr>
<tr>
<td>( ) Rarely</td>
<td></td>
</tr>
<tr>
<td>( ) Sometimes</td>
<td></td>
</tr>
<tr>
<td>( ) Frequently</td>
<td></td>
</tr>
<tr>
<td>( ) Always</td>
<td></td>
</tr>
</tbody>
</table>

13 How frequently do you hear air traffic controllers speaking at the speed below? (Click on the “play” icon on the left)

<table>
<thead>
<tr>
<th>a</th>
<th>( ) Never</th>
</tr>
</thead>
<tbody>
<tr>
<td>( ) Rarely</td>
<td></td>
</tr>
<tr>
<td>( ) Sometimes</td>
<td></td>
</tr>
<tr>
<td>( ) Frequently</td>
<td></td>
</tr>
<tr>
<td>( ) Always</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>b</th>
<th>( ) Never</th>
</tr>
</thead>
<tbody>
<tr>
<td>( ) Rarely</td>
<td></td>
</tr>
<tr>
<td>( ) Sometimes</td>
<td></td>
</tr>
<tr>
<td>( ) Frequently</td>
<td></td>
</tr>
<tr>
<td>( ) Always</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>c</th>
<th>( ) Never</th>
</tr>
</thead>
<tbody>
<tr>
<td>( ) Rarely</td>
<td></td>
</tr>
<tr>
<td>( ) Sometimes</td>
<td></td>
</tr>
<tr>
<td>( ) Frequently</td>
<td></td>
</tr>
<tr>
<td>( ) Always</td>
<td></td>
</tr>
</tbody>
</table>

http://revistas.pucsp.br/esp

DOI: https://doi.org/10.23925/2318-7115.2020v41i4a4
<table>
<thead>
<tr>
<th>d</th>
<th>( ) Never</th>
<th>( ) Rarely</th>
<th>( ) Sometimes</th>
<th>( ) Frequently</th>
<th>( ) Always</th>
</tr>
</thead>
<tbody>
<tr>
<td>* (a) speech rate of approximately 100 words per minute (b) speech rate of approximately 140 words per minute (c) speech rate of approximately 180 words per minute, and (d) speech rate of approximately 220 words per minute</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Speaking of your experience with international flights, approximately how much time did you spend listening to non-native speakers of English?</td>
<td>a) More than 60% of the time listening to NON-NATIVE SPEAKERS OF ENGLISH.</td>
<td>b) Between 40% and 60% of the time listening to NON-NATIVE SPEAKERS OF ENGLISH.</td>
<td>c) Less than 40% of the time listening to NON-NATIVE SPEAKERS OF ENGLISH.</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Who do you consider to be EASIER to understand in radiotelephony communications, a native speaker of English or a non-native speaker?</td>
<td>a) It is usually easier to understand a NATIVE SPEAKER OF ENGLISH in radiotelephony communications.</td>
<td>b) It is usually easier to understand a NON-NATIVE SPEAKER OF ENGLISH in radiotelephony communications.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Speaking of your experience with international flights, what was the highest number of pieces of information you listened to in a single transmission from an air traffic controller? For example, the following communication has 4 pieces of information: “ANAC 123, [maintain runway heading], [climb to 7,000 feet], [squawk 4437]. [Contact Miami Departure on frequency 122.4]”</td>
<td>a) Up to 3 pieces of information in each communication.</td>
<td>b) 4 pieces of information in each communication.</td>
<td>c) 5 pieces of information in each communication.</td>
<td>d) 6 pieces of information in each communication.</td>
</tr>
<tr>
<td>17</td>
<td>What makes it difficult for you to understand an air traffic controller communication? Choose all that apply.</td>
<td>A difficult accent</td>
<td>A high number of pieces of information</td>
<td>Unknown vocabulary</td>
<td>Complex grammar</td>
</tr>
<tr>
<td>18</td>
<td>Do you have experience working as an ICAO test rater?</td>
<td>( ) yes ( ) no</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Please leave any comments regarding the topics discussed in this questionnaire.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Chapter 6: Manuscript 5 – The construct of aviation English listening tests for pilots from three perspectives: Literature, policy, and perceptions of pilots and other key stakeholders

The manuscript in this chapter consists of a book chapter that has had its abstract approved by the book’s editors and publisher (Bloomsbury). The chapter will be submitted by September 1, 2023. Feedback on the chapter will be sent to the authors by November 1, 2023. The final version of the chapter is due to be sent to the editors by February 1, 2024, and the manuscript is due to be submitted to Bloomsbury by March 1, 2024. The working title of the book is Research, Assessment and Pedagogy in Global Aviation English, and its editors are Dr. Eric Friginal, from the Hong Kong Polytechnic University, Dr. Malila Prado, from the Beijing Normal University-Hong Kong Baptist University-United International College, and Jennifer Roberts, from Embry-Riddle Aeronautical University. This chapter synthesizes the results of the three previous chapters and provides the definition of the construct of an aeronautical English listening test for pilots.
The construct of an aeronautical English listening tests for pilots from three perspectives: the literature, the policy, and the perceptions of pilots and other key stakeholders

Angela C. de M. Garcia

Abstract
Since the publication of the International Civil Aviation Organization (ICAO)’s language proficiency requirements, a number of different tests have been developed and implemented around the world to assess pilots’ and air traffic controllers’ (ATCOs) proficiency in English. However, researchers have questioned the clarity and appropriateness of the policy as well as the reliability of tests (e.g., Alderson, 2011; Douglas, 2004; Emery, 2014). ICAO has recently recognized that, over the years, multiple interpretations of the policy have led to practices that might have undermined the meaningfulness of aeronautical English tests (ICAO, 2022). This chapter summarizes and complements the results of three analyses that were conducted to investigate what a test designed to assess pilots’ listening comprehension should assess. These analyses looked at the construct of a pilots’ listening test from three different perspectives: the literature, the ICAO rating scale, and the perceptions of pilots and other key stakeholders. This chapter concludes with a definition of this listening construct by combining the results of these three analyses. A better understanding of the listening construct can help minimize differences between test implementations around the world and further contribute to more standardized and meaningful testing practices.

Keywords: ICAO language proficiency requirements, language test development, aeronautical English assessment, listening construct, aviation English
1. Introduction

The International Civil Aviation Organization (ICAO) language proficiency requirements (LPRs) stipulate that pilots and air traffic controllers (ATCOs) need to “demonstrate their ability to speak and understand the language used for radiotelephony communications” (ICAO 2022, A-1). ICAO developed a rating scale to be used in the assessment of pilots and ATCOs worldwide. This rating scale is composed of six analytic criteria: pronunciation, structure, vocabulary, fluency, comprehension, and interactions and is divided into six proficiency levels (level 1, Pre-elementary; level 2, Elementary; level 3, Pre-operational; level 4, Operational; level 5, Extended; and level 6, Expert). Since the publication of the LPRs in 2003, civil aviation authorities and testing service providers have been working on the design, development, and administration of language tests that aim to assess speaking and listening. However, this was a new testing context and little was - and perhaps still is - known about the nature of the language used for radiotelephony communications. An important question to be asked then - and still today - is “what should these tests assess?” An essential step in the development of aeronautical English tests is the definition of the construct, or, in other words, a description of the abilities pilots and ATCOs need to communicate. The characteristics of the context in which this communication occurs play a critical role because test developers are interested not only in what test takers can do, but also under what circumstances (Buck 2001). The focus of this chapter is therefore on the second element of the ICAO requirements, which is to understand [emphasis added] the language used for radiotelephony communication” rather than on the first part, which is to speak.
Understanding an air traffic control communication encompasses more than just understanding English. An interesting case was described by an experienced aeronautical English examiner who gave a lecture to a group of examiners in Brazil (Cruz 2017). She showed a video of an experiment she conducted with a colleague who held a Cambridge C2 Proficiency Certificate (the highest level of qualification in English). Her colleague took a mock test of the Brazilian language proficiency test for pilots, the *Santos Dumont English Assessment* (SDEA), and, although she could understand most words from the recordings that were played, she was completely lost, and would have failed the test. A similar scenario happened on all four occasions when I played a recording of an air traffic control communication to classes of postgraduate students in Applied Linguistics. I was told by both native and non-native speakers that although English was being used, they could not understand what was going on in the recording. This confirms the importance of background knowledge in listening comprehension. As Buck (2001) advocates:

> If the listener shares the same knowledge as the speaker, much of what is being said can be understood by means of inferences based on shared background knowledge. However, if the listener has no knowledge relevant to the particular events being described in the text, then it will be more difficult to make inferences, and comprehension will be more dependent on interpreting the linguistic information. (Buck 2001, 20)

Thus, understanding pilot-ATCO communications encompasses more than just understanding everyday English. Professional background knowledge is essential to understanding this type of communication.
The purpose of this chapter is to describe the construct of pilots’ aeronautical English listening tests from three different perspectives: the literature, the policy, and the perceptions of pilots and other key stakeholders. The combination of these three perspectives may provide a unique insight into the construct by first looking at what others have already learned about a topic, understanding the policy since even if not everyone agrees with it, test developers must comply with it, and, finally, investigating key stakeholders’ views on the issue. Although the objective is to define the construct of listening tests for pilots, the findings may also be valuable for the development of listening tests for ATCOs. This chapter draws upon the findings of three analyses I conducted during my doctoral work in Applied Linguistics and Discourse Studies at Carleton University. It starts by discussing the listening construct according to the literature on language assessment (Garcia 2022). A review of the literature on what listening involves is provided along with a list of features that characterizes this specific type of listening (namely listening to aeronautical English). Then, the chapter focuses on the construct embedded in the ICAO rating scale. An analysis of the listening construct underpinned by the descriptors for comprehension is presented. Additionally, ways in which the descriptors for the other criteria in the rating scale can inform the definition of the listening construct are discussed (Garcia 2023). After that, an overview of the perceptions of pilots and other key stakeholders on the listening construct of aeronautical English tests is presented (Garcia and Fox 2020). Finally, an attempt is made to define the listening construct of aeronautical English tests by joining the results of these three analyses.

2. The aeronautical English listening construct according to the literature

Of the four traditional language skills (writing, reading, speaking, and listening), listening is the skill people use the most. According to Feyten (1991, 174), “more than forty-five percent of our
total communication time is spent in listening.” Pilots and air traffic controllers also spend much of their communication time listening to one another. Moreover, pilots seem to need to use this ability even more than ATCOs, as ATCOs’ turns are longer than pilots’ in both routine and problem transactions (Clark, Morrow, and Rodvold 1990).

### 2.1 Frameworks for the assessment of listening comprehension

Listening is a cognitive process that takes place inside the mind of the listener (Field 2019). It consists of an “intentional conversion of external stimuli into meaningful information” (Rost and Brown 2022, 238). A number of scholars have developed lists of skills and strategies involved in the listening process. In this subsection, the theoretical overviews of four influential scholars are presented: Weir’s (1993) framework for testing listening comprehension, Buck’s (2001) traditional framework for describing listening ability, and two more recent conceptualizations: Rost (2016) and Rost and Brown (2022), and finally Field (2019). These four models offer partly complementary, partly overlapping understandings of what L2 listening comprehension entails and, thus, it is worth examining each one in turn before comparing them.

Weir (1993) explains that though listening may be assessed together with speaking, there are situations in which listening occurs in isolation (e.g., listening to lectures, announcements, recorded messages, and the broadcast media in general). He argues that:

It is essential in these cases to decide on the conditions and operations that merit inclusion in a test of listening comprehension. We need to identify the situational features and important facets of interaction involved in the desired performance. It might be even more important to specify these in assessing comprehension abilities as one does not have a tangible product of speech or writing to attach behavioral labels to. To determine what
is a satisfactory performance one needs to specify in some detail what it is the candidate can do (operations), under what circumstances (conditions). (Weir 1993, 98)

Weir proposes a framework for testing listening comprehension that is divided into four categories: direct meaning comprehension, inferred meaning comprehension, contributory meaning comprehension (microlinguistic), and note taking from writing. He acknowledges that “such taxonomies of sub-skills have the status of hypothesis only, premised on what experience and opinion suggest are important” (Weir 1993, 99).

In his seminal book on the assessment of listening, Buck (2001) suggests a framework for describing listening ability. His framework is divided into language competence and strategic competence. Language competence includes grammatical, discourse, pragmatic, and sociolinguistic knowledge, whereas strategic competence includes cognitive and metacognitive strategies. For Buck, listening consists of “a top-down process in the sense that the various types of knowledge involved in understanding language are not applied in any fixed order – they can be used in any order, or even simultaneously, and they are all capable of interacting and influencing each other” (Buck 2001, 3). This perspective differs from a bottom-up view, such as that presented by Clark and Clark (1977), in which the listening process was believed to occur in a definite sequence, going gradually from the lowest level (phonemes) to the next level until the listener could understand the meaning of the utterance. Buck’s (2001) work remains a core tenet of listening assessment theory today. In fact, Buck is the external reference given by the ICAO Test Design Guidelines (ICAEA n.d.) when discussing the construct of listening.

Rost (2016) describes the nature of the listening process as composed of four types of cognitive operations: neurological, linguistic, semantic, and pragmatic processing, while Rost and Brown (2022) focus on the last three categories. They argue that the linguistic processing
consists of bottom-up operations and that “for fluent L1 and L2 listeners, most gaps and inaccuracies in this bottom-up linguistic processing can be compensated for through ‘top-down’ semantic and pragmatic processes that are less dependent on accurate perception of the incoming signal” (Rost and Brown 2022, 238).

Field (2013; 2019) lists similar knowledges and competencies. He proposes a model of listening that describes the operations performed by expert listeners, which, in his view, should be the goal of second language learners. Field believes that listening involves five different operations: input decoding, lexical search, parsing, meaning construction, and discourse construction. Field (2013; 2019) also believes that for listening to take place, both bottom-up and top-down processes must occur simultaneously. For him, the performance of second language listeners can be affected by two general factors: lack of linguistic and cultural knowledge, and lack of expertise in handling connected speech and in using strategies. Table 1 provides an overview of what listening involves according to the four mentioned models.
Table 1. What listening involves according to the literature

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Direct meaning comprehension</strong></td>
<td><strong>Language competence</strong></td>
<td><strong>Neurological processing</strong></td>
<td><strong>Lower level processes</strong></td>
<td></td>
</tr>
<tr>
<td>Listening for gist</td>
<td>Grammatical knowledge</td>
<td>Hearing</td>
<td>Input decoding</td>
<td></td>
</tr>
<tr>
<td>Listening for main ideas</td>
<td>Phonology</td>
<td>Consciousness</td>
<td>Lexical search</td>
<td></td>
</tr>
<tr>
<td>Listening for specifics</td>
<td>Stress, intonation</td>
<td>Attention</td>
<td>Parsing</td>
<td></td>
</tr>
<tr>
<td>Determining speaker’s attitude or intentions toward listener or topic</td>
<td>Spoken vocabulary</td>
<td><strong>Linguistic processing</strong></td>
<td>Higher level processes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Spoken syntax</td>
<td>Phonological processing</td>
<td><strong>Meaning construction</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Discourse knowledge</td>
<td>Segmental processes</td>
<td>Identifying speaker intention and context</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pragmatic knowledge</td>
<td>Supra-segmental processes</td>
<td><strong>Inference</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sociolinguistic knowledge</td>
<td>Lexical processing (word recognition)</td>
<td><strong>Reference</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Inferred meaning comprehension</strong></td>
<td><strong>Strategic competence</strong></td>
<td><strong>Syntactic processing</strong></td>
<td><strong>Discourse construction</strong></td>
<td></td>
</tr>
<tr>
<td>Making inferences and deductions</td>
<td>Cognitive strategies</td>
<td>Knowledge activation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relating utterance to the social and situational context</td>
<td>Comprehension processes</td>
<td>Comprehension building</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recognizing the communicative function of utterances</td>
<td>Storing and memory processes</td>
<td><strong>Semantic processing</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deducing meaning of unfamiliar lexical items from context</td>
<td>Using and retrieval processes</td>
<td>Knowledge activation</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Contributory meaning comprehension (microlinguistic)</strong></td>
<td>Metacognitive strategies</td>
<td>Word formation (polysemic)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Understanding phonological features (stress, intonation, etc.), concepts, discourse markers, syntactic structure, grammatical and lexical cohesion, lexis</td>
<td>Assessing the situation</td>
<td><strong>Pragmatic processing</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Listening and writing</strong></td>
<td>Monitoring</td>
<td>Situational framing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ability to extract key points</td>
<td>Self-evaluating</td>
<td>Inferring speaker intention</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Self-testing</td>
<td>Communication strategies</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Although these categories and structure of what is involved in listening differ, they list similar knowledge, competencies, and processes. What Weir (1993) calls “contributory meaning
comprehension,” for example, Buck (2001) calls “grammatical knowledge,” Field (2019) calls “language-related knowledge,” and Rost (2016) calls “linguistic processing.” In addition, other relationships can be established across the four lists. Simply put, we may conclude that listening involves linguistic and background knowledge, as well as neurological and pragmatic processing. It is important to note that, when assessing linguistic knowledge, listening test developers, as Buck (2001) points out, should aim to assess test takers’ procedural knowledge rather than declarative knowledge (assess test takers’ ability to understand the language instead of their knowledge about the language per se). Figure 1 presents a summary of what is involved in listening according to the literature.

![Figure 1. What is involved in listening according to the literature](image)

2.2 Listening to aeronautical radiotelephony communications

According to Green (2017), “the definition of what listening is will differ according to the purpose of the test and also the target test population” (Green 2017, 29). Listening to aeronautical radiotelephony communications is not the same as listening to everyday language. It is a highly specific form of discourse. ICAO requires pilots and ATCOs to use standardized
phraseology to communicate. The ICAO standardized phraseology is a kind of formulaic speech created specifically for ATCOs and pilots. The *Manual on the Implementation of ICAO Language Proficiency Requirements* (DOC 9835) defines formulaic speech as “a restricted or coded use of language comprising fixed standard phrases or lexical and syntactical routines, developed either by consensus for highly repetitive communications (e.g., everyday exchanges of greetings) or formally prescribed for special or professional purposes” (ICAO 2010, ix).

Whenever there is no standardized phraseology for a particular situation, plain language should be used. Plain language is defined in DOC 9835 as “the spontaneous, creative and non-coded use of a given natural language” (p. x). If plain language is needed, the recommendation is that messages must be as clear, concise, and unambiguous as possible.

One of the differences between listening to radiotelephony communications and other types of listening is the specificity of the topic, which is always related to air traffic control and to the operation of aircraft (Garcia 2022). Another unique characteristic of this kind of listening is that it occurs through the radio, with no overlapping. In this context, speakers may request clarifications, if necessary. Figure 2 shows the features of aeronautical English listening that may contribute to communication problems.

![Figure 2. Features of aeronautical English listening that may contribute to miscommunication](image-url)
These characteristics are interrelated and may influence each other. For example, if workload is heavy, pilots tend to speak at a faster speech rate (Prinzo and Britton 1993). Moreover, some of these characteristics have a greater impact on communication than others. According to Estival and Molesworth (2016), for instance, ATCOs’ transmissions with an increased number of items, as well as pilots’ heavy workload, may impact comprehension more than ATCOs’ speech rate. These features are briefly discussed below:

- **Absence of visual support** – Communication takes place via the radio, with nonverbal communication, which helps listeners in face-to-face interactions (e.g., facial expressions, body language, and gestures), not present.

- **Wide range of accents** – As pilots fly all over the world and ATCOs control flights from anywhere in the world, pilots and ATCOs are exposed to a wide variety of accents from both native and non-native speakers of English.

- **Wide range of aviation-related topics** – Although the topics of radiotelephony communications are always related to the aviation context, there is a wide range of such topics, especially when dealing with abnormal and emergency situations, including problems related to technical operations, weather, health, and passenger behavior (see ICAO 2010, Appendix B).

- **Stressful work environment** – Stress caused by the need to deal with challenging situations may contribute to an increase in failures to understand communications (Dismukes, Kochan, and Goldsmith 2018).
- **Increased mental workload** – Multitasking and communicating at the same time as flying may adversely affect communication accuracy (Estival and Molesworth 2016).

- **Poor quality of transmission** – As would be expected, technical issues, such as radio interference and cut-off transmissions, interfere with ease of understanding (Prinzo and Thompson 2009). However, with advances in technology, this problem seems to be less serious today than in the past (Garcia and Fox 2020).

- **Frequency congestion** – Research suggests that radio frequency congestion (the frequency being shared by a high number of operators) is one of the main causes of miscommunication (see Prinzo and Britton 1993). However, Estival and Molesworth’s (2016) study indicates that radio frequency congestion does not appear to affect pilots’ ability to understand communications.

- **Interference from intercultural factors** – Intercultural factors related to power distance, face-work strategies, conflict management, communication styles, and (non)collaborative behavior appear to have an impact on the efficiency of radiotelephony communications (Monteiro 2019).

- **Noisy environment** – Aircraft cockpits are considerably noisier than a typical office. Language comprehension is highly vulnerable to this noise, especially in the case of non-native speakers (Molesworth 2016).

- **Fast speech rate of speaker** – According to ICAO recommendations, pilots and ATCOs should speak at a rate of 100 words per minute, or a little faster “if the situation requires” (ICAO 2010, 4-12). However, research shows that pilots and ATCOs frequently speak considerably faster than that recommendation (Prinzo and Britton 1993; Garcia and Fox 2020).
2020). Research also indicates that “increased ATC [air traffic control] speech rate is particularly problematic for EL2 [English as a second language] pilots with low pilot qualifications” (Estival and Molesworth 2016, 247).

- *Long transmissions* – Pilots and ATCOs’ transmissions should be concise. However, single transmissions sometimes include a high number of items, which may contribute to an increase in communication failure (Estival and Molesworth 2016). This likely occurs due to capacity limitations in working memory (Morrow and Rodvold 1993). Barshi and Farris (2013) argue that “when both the pilot and the controller share the same native language, controller delivery of messages of three commands may provide an optimal balance of the accuracy and efficiency constraints inherent in the controller-pilot communication environment” (Barshi and Farris 2013, 122).

3. The listening construct in the ICAO rating scale

Whenever test developers need to create a test based on a test purpose and mandate dictated by a policy, they need a clear understanding of this policy. In the case of aeronautical English assessment, the policy was stipulated by ICAO and implemented by member states. The ICAO rating scale was developed by the Proficiency Requirements in Common English Study Group (PRICESG), which consisted of “Contracting States and international organizations, operational and linguistic experts with backgrounds in aviation (pilots, air traffic controllers and civil aviation authority representatives), aviation English training and applied linguistics” (ICAO 2010, 1-3). The rating scale was published in 2003 and is still used in the assessment of pilots and ATCOs worldwide.

An experienced test developer and rater who participated in Garcia’s (2015) study pointed out that “there are six criteria in the scale … 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%.” (Garcia 2015, 101). DOC 9835 concedes that “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). Although a number of researchers have argued that the ICAO policy and its rating scale should be revised (e.g., Douglas 2014; Garcia 2015; Knoch 2009), they remain unrevised and in use.

In addition to Annex 1 and DOC 9835, ICAO is soon going to publish a handbook containing test design guidelines (ICAEA n.d.). According to these guidelines, “test instruments need to contain test tasks dedicated to assessing listening comprehension separate from tasks designed to assess speaking performance” (ICAEA, n.d., Criterion 3). In response, the purpose of Garcia’s (2023) study was to investigate the listening in isolation construct underpinning the comprehension descriptors in the ICAO rating scale, and to investigate how the other descriptors could help inform this listening construct. However, it should be highlighted that, “although it is necessary to assess listening in isolation, the assessment of interactive listening is also essential in this context as most of the listening performed by pilots and ATCOs happens as part of an interactive conversation between them” (Garcia 2023, 71).

In her content analysis of the rating scale, Garcia (2023) concluded that, according to the ICAO rating scale, listening tests should assess pilots and ATCOs’ ability to:

- Understand common, concrete, and work-related topics – Although it is not clear what is meant by common, concrete, and work-related topics (Douglas 2004; Knoch 2009), the descriptors for vocabulary may help to understand it as they also refer to common, concrete, and work-related topics (levels 4 and 5). As the vocabulary descriptors also mention idiomatic vocabulary (levels 5 and 6), a wide variety of familiar and unfamiliar
topics (level 6), and nuanced vocabulary (level 6), we may conclude that common, concrete, and work-related topics probably do not include idiomatic, nuanced, or unfamiliar vocabulary. These topics are likely those that are familiar to pilots and ATCOs, related to their work, frequently used (common), and concrete (more tangible? Not nuanced?). The need to assess lexical knowledge in listening tests is upheld by Rost and Brown (2022), who argue that “for L2 listeners, lexical knowledge is the most significant variable contributing to listening proficiency, so in order to become more skilled, it is important for L2 learners to be able to recognize words they do know, and to increase their ‘listening lexis’ — words they can readily recognize in spoken input” (Rost and Brown 2022, 241).

- Understand accents or varieties sufficiently intelligible for an international community of users (level 4) as well as a range of speech varieties (dialect and/or accent) or registers (level 5) – It is not clear what accents or varieties may be considered sufficiently intelligible for an aeronautical community of users. However, the descriptors for pronunciation may shed some light on this. According to the pronunciation descriptors, the elements that may contribute to intelligibility are both the segmental and the suprasegmental aspects of pronunciation (“pronunciation, stress, rhythm, and intonation”). These elements of pronunciation may or may not be influenced by the speaker’s first language or regional variation, as the level 6 descriptors acknowledge that pronunciation, stress, rhythm, and intonation may be “possibly influenced by the first language or regional variation” (ICAO 2010, A-7). Thus, an aeronautical English listening test could include input with different pronunciation (individual sound segments), stress, rhythm,
and intonation. The ability to understand a broad range of accents should also be mentioned in the level 4 descriptors (Kim and Billington 2016).

- *Understand linguistic or situational complications or an unexpected turn of events* – A number of factors, such as an unfamiliar accent, topic, or grammar structure, or a long input, may be considered a linguistic complication. Factors related to the fluency descriptors, such as a fast speech rate, “inappropriate phrasing and pausing, slowness in producing language, use of too many fillers, or use of inappropriate discourse markers or connectors” (Garcia 2023, 81) may also contribute to a linguistic complication. Situational complications may include situations that are uncommon, such as abnormal, urgent, and emergency situations as well as situations with which the test takers are not familiar. An unexpected turn of events, which the descriptors list as another possibility, may also be considered a situational complication, caused, for example, by a misunderstanding (mentioned in the level 4 descriptors for interactions), “or by a response which is not immediate, appropriate, or informative, among other possibilities” (Garcia 2023, 81).

- *Ask for clarification when comprehension fails* – In addition to the skills listed above, aeronautical English listening tests should also assess test takers’ ability to use clarification strategies. The level 4 descriptors for interactions also include the ability to check for information. Although this ability is only mentioned in the descriptors for level 4, it should also apply to levels 5 and 6 (Garcia 2015).

Since accents may be considered a linguistic complication, and an unexpected turn of events may be considered a situational complication, the construct of listening in the ICAO rating scale is summarized in Figure 3.
Additionally, tests that aim to assess pilots and ATCOs’ at level 6 (expert) should include the following:

- **Comprehension in nearly all contexts** – The descriptors for level 6 in the comprehension and interactions criteria mention, respectively, “comprehension … in nearly all contexts” and the ability to interact “with ease in nearly all situations.” These descriptors are very vague. What exactly is meant by “nearly all contexts or situations”? Does this include contexts unrelated to aviation? If so, what is its relevance? Even if all context were relevant, how could a test reflect them all? In addition, the descriptors imply that there are contexts in which a level 6 pilot or ATCO might not have a good understanding or might have difficulty interacting (Garcia 2015). For these reasons, I argue that these descriptors should not inform the construct definition.
• Comprehension of linguistic and cultural subtleties – Comprehension of linguistic subtleties may include comprehension of nuanced vocabulary, which is mentioned in the vocabulary descriptors for level 6 (e.g., nuances of meaning between words, or collocations [Pearson, Hiebert, and Kamil 2007]). Though comprehension of cultural subtleties is likely to be important to all pilots and ATCOs, this item only appears in the level 6 descriptors for comprehension (Garcia 2015). Monteiro (2019) also highlights that comprehension of cultural aspects enhances understanding between multi-cultural pilots and ATCOs. One of the key components of Monteiro’s matrix of construct specification within the aviation radiotelephony domain is the combination of intercultural awareness and competence.

4. The aeronautical English listening construct and perceptions of key stakeholders

So far, we have discussed the construct of listening according to the literature and the ICAO rating scale. It is now important to turn our attention to the test takers. When developing a test to assess a language in a professional setting, test developers should take into consideration what criteria members of that professional culture consider important to be assessed by the test. As Jacoby and McNamara (1999) argue, “since members of professional cultures are faced with the interactional problem of articulating to one another the grounds for their own evaluations of performance and recommendations for improvement, an analysis of ‘indigenous assessment’ may provide a window onto an insider’s view of the complex issues involved in communicating competently in some particular domain” (p. 214). Thus, this section reports the results of Garcia and Fox’s (2020) study, which aimed to investigate the perceptions of pilots and other key stakeholders regarding the construct of aeronautical English listening tests for pilots. In this study, 156 airplane and helicopter pilots (mostly from Brazil) responded to an online
questionnaire on what pilots’ listening tests should assess, and six key stakeholders (language proficiency raters, aviation English researchers, and an ATCO) were interviewed in order to help interpret the questionnaire's findings. Among other topics, questions addressed the features of aeronautical English listening shown in Figure 2. Results show that there is a need to assess pilots’ ability to:

- **Recognize words and numbers** – The vast majority of survey participants (98.7%) believed that it is important or extremely important to assess the ability to recognize words and numbers. As one interviewee explained, this is extremely important because there is a great deal of numerical information in aviation, including “runway numbers, altitudes, speed, directions, and headings” (Garcia and Fox 2020, 22). Understanding words is also considered important, especially words specific to the aviation context.

- **Understand main ideas** – 98.1% of participants also acknowledged that it is either important or extremely important to assess the ability to understand main ideas.

- **Understand both plain English and phraseology** – Although ICAO only requires plain English to be formally assessed (ICAO 2010), the results of Garcia and Fox’s (2020) study indicate that both plain English and phraseology should be assessed. In fact, 94.9% of the pilots who participated in the study either agree or strongly agree that phraseology should be assessed, a higher percentage than for plain English, which was 85.7%.

- **Extract specific meaning from both short and long transmissions** – 89% of participants reported that it is either important or extremely important to assess the ability to extract specific meaning from both short and long transmissions. One of the participants who were interviewed explained that it is important to identify what information is crucial and what pieces of information are not so important.
• **Understand speech at different rates of delivery** – The pilots who participated in the study listened to four air traffic control recordings in which the same script was recorded by the same person at four different speech rates (100, 140, 180, and 220 words per minute). They were asked how frequently they were likely to hear ATCOs speaking at each of these four speech rates. According to the pilots who participated in the study, the most frequent ATCOs’ speed of delivery seems to be 180 words per minute, which upholds Prinzo and Britton’s (1993) finding that ATCOs speak much faster than the recommended 100 words per minute. Five of the six participants who were interviewed also believe that ATCOs speak faster than ICAO’s recommendation. According to the sixth participant, only ATCOs from the United States speak faster. The results show that ATCOs speak at a range of different speech rates, which may be the reason why the majority of participants (79.5%) reported that it is either important or extremely important to assess pilots’ ability to understand speech at different speeds.

• **Understand different accents** – Participants reported that the frequency at which pilots hear unintelligible pronunciation in radiotelephony communication is between “sometimes” and “frequently.” Two participants who were interviewed explained that unintelligible pronunciation does not necessarily mean pronunciation by a non-native speaker. In fact, they reported that it is sometimes easier to understand a non-native speaker of English than a native speaker. According to one of the interviewees, what influences ease of understanding may be familiarity with the speaker’s accent. Again, most participants (72.4%) reported that it is either important or extremely important to assess pilots’ ability to understand different accents. This upholds Kim and Billington’s
argument that the assessment of a range of accents should be included in the comprehension rating scale descriptors for level 4.

- **Understand transmissions with more than three items** – In spite of Barshi and Farris’s (2013) recommendation that radiotelephony messages should contain a maximum of three items per transmission, pilots who answered Garcia and Fox’s (2020) survey reported that ATCO’s transmissions sometimes include more than this number. Participants explained that the number of items depends on the phase of the flight. One interviewee said that “there might be up to 20 elements in a single transmission, but there are typically around three or four” (p. 18). Participants also pointed out that in case a long transmission is not fully understood, the problem may be due to memory limitations, which would uphold Morrow and Rodvold’s (1993) results.

- **Understand basic grammar** - It is important to assess basic grammar. As one of the participants who were interviewed said, pilots “need to understand the basic verb tenses like present simple, present continuous, past, and future” (Garcia and Fox 2020, 21).

Other themes that were recurrent in the interviews and that should be mentioned are the assessment of pilots’ ability to understand transmissions in spite of background noise, during increased workload, and in congested radio frequencies, to understand messages that differ from what was expected, and to use strategies in order to understand.

The assessment of other abilities seems to be less important than those already mentioned. However, they should be included to some extent (e.g., be still part of the test construct but with a lower weight and be addressed in tests that aim to assess pilots at level 6). These secondary abilities include being able to:
- **Understand the meaning of uncommon words and expressions** – Pilots who participated in the study reported that they “rarely” or “sometimes” hear words or expressions they do not understand. Although this does not occur frequently, 50% of them consider it either important or extremely important to assess the ability to understand the meaning of uncommon words and expressions. Four participants who were interviewed also believe that this is important because pilots may experience unusual situations. However, one of them highlighted that the word or expression should still be related to aviation. The other two interviewees argued that this is not so important because pilots can ask for clarification when they do not understand the meaning of a word. This is likely why fewer questionnaire respondents consider this aspect to be important. Regarding the use of idiomatic expressions, the six participants who were interviewed reported that ATCOs use idioms, especially, as three of them stated, in the United States.

- **Recognize indirect meaning** – The need to recognize implied ATCO’s intentions or attitudes and (un)professional behavior (e.g., politeness, sarcasm, professionalism, and anger) was considered important but only “somewhat” by the survey participants. Half of the interviewed participants consider this an important aspect to be assessed. One of them explained that ATCOs are not usually unprofessional. Although one participant argued that the ability to recognize indirect meaning should be assessed at all levels, two other participants argued this should be included only in tests designed to assess level 6 pilots. Another participant recognized the importance of at least including this ability in pilots’ training.

- **Understand complex grammar** – The results showed that it may not be very important to assess complex grammar. The frequency at which pilots reported hearing sentences they
cannot understand because of structure was between “rarely” and “sometimes.” Most interviewees believe that this is not important. However, tests developed to assess expert level 6 should probably include the ability to understand complex grammar.

- Understand low quality transmissions – Participants acknowledge that low quality transmission may impact ease of understanding. However, as two interviewees pointed out, this seems not to be as much of a problem nowadays as it used to be in the past.

5. A description of the aeronautical English listening construct from three perspectives

Having explored the construct of listening from the perspective of the literature, the ICAO rating scale, and the perceptions of pilots and other key stakeholders, it is time to put these three perspectives together in order to define the construct of aeronautical English listening tests. Figure 4 shows a summary of the listening construct according to the three perspectives discussed in this chapter. The constructs shown in Figure 4 should be used to create tasks that aim to assess pilots’ listening comprehension at levels 4 and 5.
Figure 4. Summary of the aeronautical English listening construct from three perspectives
Figure 5 shows abilities that should be included to a lesser extent in a pilots’ listening test. The inclusion of tasks that aim to assess these abilities may help to discriminate between level 6 and lower test takers.

![Diagram showing Policy and Pilots' perceptions with subcategories]

Understand linguistic subtleties
Understand cultural subtleties

Understand the meaning of uncommon words and expressions
Recognize ATCO’s attitude and intentions
Understand complex grammar
Understand low quality transmissions

Figure 5. Abilities that should be assessed to a lesser extent

Finally, Table 2 shows the definition of the construct of an aeronautical English listening test for pilots. The conceptual foundation for this construct definition is the sociocognitive theory proposed by Weir (2005), which takes into account both the cognitive processes and the features of the context. This construct definition follows Weir’s (1993) suggestion to specify what test takers “can do (operations), under what circumstances (conditions)” (Weir 1993, 98). It is important to remember that some operations and conditions listed in Table 2 are more relevant than others, as discussed in this chapter.
Table 2. Definition of the construct of aeronautical English listening test for pilots

<table>
<thead>
<tr>
<th>Operations (can do)</th>
<th>Conditions (under what circumstances)</th>
</tr>
</thead>
<tbody>
<tr>
<td>● Recognize common, concrete, aviation-related words and expressions (understanding numbers and what they refer to is crucial),</td>
<td>● Speech at different rates of delivery (from 100 to 220 words per minute)</td>
</tr>
<tr>
<td>● Understand uncommon, unfamiliar words and expressions (including nuanced vocabulary, idioms, and collocations)</td>
<td>● Different accents and registers (including when pronunciation, stress, rhythm or intonation interfere with ease of understanding)</td>
</tr>
<tr>
<td>● Understand both plain English and phraseology</td>
<td>● Short and long transmissions (including transmissions with more than three items)</td>
</tr>
<tr>
<td>● Understand main ideas</td>
<td>● Transmissions with lack of orality features (e.g., redundancy, inappropriate phrasing and pausing, slowness, with many fillers, without appropriate discourse markers or connectors)</td>
</tr>
<tr>
<td>● Extract specific meaning</td>
<td>● With background noise</td>
</tr>
<tr>
<td>● Understand basic and complex grammar</td>
<td>● During increased workload</td>
</tr>
<tr>
<td>● Understand messages that differ from what was expected</td>
<td>● In congested radio frequencies</td>
</tr>
<tr>
<td>● Identify misunderstandings</td>
<td>● Transmissions with low quality</td>
</tr>
<tr>
<td>● Use strategies in order to understand (e.g., ask for clarification when comprehension fails)</td>
<td></td>
</tr>
<tr>
<td>● Understand cultural subtleties / recognize ATCOs’ attitude and intentions (e.g., politeness markers, sarcasm, professionalism, anger)</td>
<td></td>
</tr>
</tbody>
</table>

6. Conclusion

The goal of this chapter was to provide unique insights into the test construct of pilots’ listening comprehension by combining three lenses: the literature, the policy, and the perceptions of pilots and other key stakeholders. While this construct definition will assist in the
development and improvement of listening tests for pilots, it should also be useful for the assessment of ATCOs. With this construct definition, test developers can proceed to the next step in the test development process, which is writing the test specifications.

Although, as Taylor and Geranpayeh (2011) point out, “lists of sub-skills can be difficult to operationalize systematically in the form of test instruments” (p. 92), having a clear understanding of the construct to be measured by a test is paramount and is at the very heart of any language assessment process. If the construct is not clearly defined, chances are that the test is not testing what it is supposed to test, a validity problem, as the inferences that are made based on test scores may not be considered valid. As Monteiro and Fox (2022) point out, “there is clearly a great deal at stake for those involved in RT [radiotelephony] communications and for all who are directly or indirectly affected by the decisions made on the basis of test scores” (Monteiro and Fox 2020, 165). Thus, test developers need to be accountable to stakeholders for the decisions made based on test scores. In the case of aeronautical English assessment, stakeholders include pilots and ATCOs as test takers, airlines as employers, passengers, and other persons on the ground who may be affected by the consequences of language testing.
References


Cruz, Silvana. 2017 (September 26). “Listening Comprehension in Aviation English Assessments.” Lecture delivered to the Santos Dumont English Assessment Annual Recurrent Training Course for Raters, São José dos Campo, Brazil).


Estival, Dominique, and Brett Molesworth. 2016. “Native English Speakers and EL2 Pilots: An Experimental Study.” In *Aviation English: A Lingua Franca for Pilots and Air Traffic Control*.


https://doi.org/10.22190/JTESAP230220007G

Garcia, Angela C. M., and Janna Fox. 2020. Contexts and constructs: Implications for the testing of listening in pilots’ communication with air traffic controllers. The ESPrincipal, 41(4), 1-33. https://doi.org/10.23925/2318-7115.2020v41i4a4


https://doi.org/10.1093/applin/amv075


Chapter 7: Conclusion

7.1 Final approach

The language proficiency requirements for pilots and ATCOs were introduced by ICAO in an effort to increase flight safety. Although it has been twenty years since the publication of the ICAO LPRs, a great deal remains to be known about the language pilots and ATCOs use to communicate, and how to go about assessing it. Taking into consideration that:

1) ICAO emphasizes the importance of assessing listening in isolation.

2) It is indeed necessary to assess listening in isolation because:
   a. of construct-irrelevant variance;
   b. pilots need to listen to recordings such as ATIS information; and
   c. Pilots need to monitor radiotelephony communications in order to maintain situational awareness.

The purpose of this dissertation was to investigate what a test that aims to assess pilots’ listening in isolation should assess. Thus, the overarching research question which this dissertation addressed was:

*How can the aeronautical English construct be defined in a test that aims to assess pilots’ listening in isolation?*

To address this research question, three angles were adopted:

1) I conducted secondary research to look at what the literature reports about the assessment of listening and the nature of aeronautical English (Garcia, 2022).

2) I explored the aeronautical English construct underpinning the comprehension descriptors of the ICAO rating scale and how the other descriptors in the rating scale could inform the definition of the listening construct (Garcia, 2023).
3) I jointly conducted an empirical investigation to explore pilots’ and other key stakeholders’ perceptions on the nature of the listening tasks performed by pilots (Garcia & Fox, 2020).

The results of these three analyses were merged in order to define the listening construct, which is the main contribution of this dissertation, as seen in Chapter 6. This construct definition follows Weir’s (1993) recommendations by specifying what test takers should be able to do and under what circumstances (see Table 2, Chapter 6). The results of the research presented in this dissertation add to the existing body of knowledge and fill a research gap regarding what aeronautical English listening tests should assess by providing a definition of the listening construct.

7.2 Limitations

Although care has been taken in planning and conducting the research and analyzing the results presented in this dissertation, some limitations should be identified. First, the scope of my dissertation was limited to the analyses of the literature, the rating scale, and the perceptions of pilots and other key stakeholders. Avenues for further research could broaden this scope by, as proposed in the project described in Chapter 2, looking at the stated, perceived, and operationalized constructs of existing tests. Also, the final construct definition could be presented to a group of experts in order to refine and validate it. A number of approaches could have been taken to define the test construct, and the approach I took in this dissertation represents only one of the possible options.

Another limitation is related to the literature review. Although I believe to have covered most of the relevant literature related to the listening construct and to the assessment of aeronautical English, it is not possible to review all existing literature given the exponential
growth of publications. Also, as this is a manuscript-based dissertation, once the manuscript has already been published, it is not possible to update it in terms of more recent literature. For a comprehensive review of the listening construct, readers may refer to the recently published article by Aryadoust and Luo (2023), in which they systematically review 157 papers on the definition and operationalization of the listening construct.

Importantly, all research is inherently subjective. The research results may have been influenced by my subjective and perhaps unconsciously biased interpretation of the qualitative data. Moreover, the results of Chapter 5, in which there were human participants, reflect the varied and subjective interpretations of the issues by the research participants. Furthermore, these results are based on what participants claimed to think, which may be different from reality. Moreover, the participants were mostly from Brazil. Thus, these results may not reflect the perceptions of a wider community of aeronautical English users.

7.3 Next flights?

Following the construct definition, the next step in the test development process is writing the test specifications, and this is what I plan to be doing at ANAC once I go back. The defined construct along with other information (e.g., a description of the purpose of the test, of the test takers, and of the target domain) can form a document called the design statement, which will guide test developers in the next phases of test development (Knoch & Macqueen, 2020). To help inform the test specifications, once the design statement is ready, test developers may produce a table of domain tasks to serve as tasks for consideration in the test, as suggested by Knoch and Macqueen. This table should include key information about the target domain tasks. Test developers may then proceed to examine if these real-world tasks are suitable to be used in a test. They would need to see whether the tasks are usable in a testing context, if they align with
the policy and requirements of the test, if they are feasible to produce within the available resources, and what might be their impact on teaching and test preparation. Knoch and Macqueen argue that real-world tasks are rarely adopted in their existing form and usually need adaptations before inclusion in a test. They explain that “test developers are therefore often required to make considerable changes to the types of tasks encountered in the domain and create tasks that only approximate the domain to some extent” (p. 100).

An important consideration when designing test tasks for a professional purpose, such as an aeronautical English test, is authenticity. Bachman and Palmer (1996) define authenticity as “the degree of correspondence of the characteristics of a given language test task to the features of a TLU [target language use] task” (p. 23). Authenticity does not mean that the tasks need to be genuinely taken from the real-world, but that tasks should be authentically designed to represent the TLU domain, and that they should be delivered and assessed authentically. The test should also enable test-takers to interact with the tasks authentically (Douglas, 2001). According to Wagner and Ockey (2023), real-world listening ability includes interactive speaking and listening (which is not the focus of this dissertation but is a crucial part of the aeronautical English construct), different accents and dialect varieties (also an important part of the aeronautical English construct, as seen in Chapter 5), as well as rapid or connected speech, and audiovisual input. Although a pilot does not see the ATCO (thus facial expressions, lip movement, gestures, etc., which usually support understanding, are not contributing factors), visual input may be used to provide context or to show what the pilot may be looking at while listening, and the ATCO may also be looking at while speaking (e.g., a photo of the airport parking zone or airfield topography, a video of aircraft taxi movement). To enhance authenticity, test developers should consider using unscripted texts, which should include spoken language
features, such as false starts and hesitations (e.g., repetitions, corrections, filled or unfilled pauses). However, although test developers need to make an effort to create authentic test tasks, the fact that a test is a simulation, not a real-life situation, inevitably affects authenticity. Thus, authenticity is a matter of degree. Corpus linguistics can be highly useful in the creation of authentic test tasks (Cushing, 2017). Aeronautical English test developers can use aeronautical English corpora to create test tasks. Pacheco et al. (2020) give practical examples of how to use corpora to create test tasks in an aeronautical English context.

Additionally, the definition of the listening construct developed in this dissertation may be useful in validation studies as test developers may use it to help develop their validity argument. To explain the meaning of test scores, inferences about the construct can be made. Test developers may use Knoch and Macqueen’s (2020) set of warrants and assumptions from the domain description inference in their LAPP validation framework.

7.4 Suggestions of further research

Suggestions for further research were given in Chapters 2, 4, and 5. To avoid repetition, I now emphasize the need to review the policy and the ICAO rating scale and add a number of recommendations to those already suggested.

The ICAO policy and its rating scale form the foundation for aeronautical English test development. As McNamara et al. (2019) argue, “if the policy is misguided, what is the point of the effort?” (p. 195). Although a number of researchers have argued that the policy and the ICAO rating scale needs to be revised (e.g., Aragão, 2018; Douglas, 2004, 2014; Garcia, 2015; Kim, 2012; Knoch 2009, 2014), more research is needed to inform the development of a new rating scale, or two rating scales: one for the assessment of speaking and another one for the assessment of listening, as suggested by Garcia (2015). Researchers may also consider other
options. After watching a presentation about the speaking construct at the Language Testing Research Colloquium (LTRC) in June of 2023 (Jong, 2023), I realized how similar the speaking and listening constructs are. In essence, the components are the same, but on the listening side, the focus is on understanding language, while on the speaking side, the focus is on producing language. Thus, ICAO LPRs reviewers could develop two rating scales, instead of keeping one rating scale that aims to assess both speaking and listening at the same time: one interpersonal rating scale to assess interactive speaking and interactive listening, and one interpretative rating scale to assess independent listening, or listening in isolation (Cutshall, 2012). To that end, researchers may follow the framework for the development of rating scales proposed by Knoch et al. (2021).

Since there is still a great deal to learn regarding the nature of aeronautical English communications and about its assessment, language testing researchers may use corpus linguistics, as Cushing (2017) suggests, to compare a specialized corpus with a general English corpus in order to “describe in greater detail the distinguishing features of language use in a particular situation” (p. 442). Thus, researchers could compare an aeronautical English corpus with a general English corpus to better understand, for example, what plain English in aeronautical communication entails.

In addition, more research is necessary to investigate the impact of technology on radiotelephony communications and its advantages and disadvantages in the assessment of aeronautical English. With advancements of data link, where pilots and ATCOs communicate through text messages, assessing reading and writing skills will likely be required (ICAO, 2016). Moreover, the development of artificial intelligence (AI) will likely impact pilot/ATCOs
communications (Hill, 2023), and assessment practices (Hart-Davidson et al., 2023). Thus, research related to advances in technology should be encouraged.

Research also needs to be conducted to define the most appropriate procedures to be undertaken in a listening test for pilots. One decision test developers need to make is to define whether test takers will be able to take notes during the listening test. As pilots take notes in the cockpit while listening to ATC communications, they should also be able to do so during a listening test. However, test developers need to define the most suitable note-taking procedures. What note-taking procedures would better reflect a real-world scenario? Once the items are being trialed, researchers may also investigate the quality of test takers’ notes and their impact on scores. Another practical decision test developers will need to make is to decide how much context should be provided for test takers. Should test takers be allowed to preview item stems and/or options in computer-based tests? If so, how much time should be given? If not, would test takers be allowed to read the items while the recording is being played or after it has finished? It would also be useful to investigate the impact of any preview (or no preview) on difficulty and scoring, since research has shown that a preview may decrease difficulty and jeopardize cognitive validity (see Yeager & Meyer, 2022).

7.5 Landing

Going back to the biblical rendition of the Shibboleth test, if the construct of a test is not clearly defined, it might be “killing the wrong people.” I would like to emphasize that the language testing of pilots and ATCOs involves very high stakes. Lives are not only metaphorically speaking, but also literally, at stake. If a construct introduces irrelevant features, or fails to consider important dimensions, we cannot trust that the inferences that are made based on test scores are valid.
I trust this dissertation has contributed at least in part to a better understanding of the assessment needs in the aeronautical communication context, and to the promotion of more consistent interpretations of the aeronautical English construct. I also hope to have encouraged more practitioners in the field of aeronautical English testing to get involved in language assessment research and to work towards the advancement of our discipline. The ultimate goal of this research was to better understand the listening construct in the context of aeronautical English and, consequently, help to increase flight safety by contributing to fairer and more meaningful assessment practices around the world.
References


Garcia, A. C. M. (2015). *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.* Master’s dissertation, Lancaster University.


technology.com/news/russian-conglomerate-develops-ai-communications-for-fifth-gen-aircraft


https://www.icaea.aero/webinars/webinars-2020


https://doi.org/10.1080/10904018.2022.2029705
Appendix A: Co-author Permission Letter

May 25, 2023

To whom it may concern,

Pursuant to article 12.4, section C of Carleton University’s Graduate Calendar, I hereby give Angela Carolina de Mornes Garcia permission to use the published work entitled *Contexts and constructs: Implications for the testing of listening in pilots’ communication with air traffic controllers* as a chapter in her dissertation. I confirm that I contributed to the chapter as a co-author and assisted in the analysis of data and writing of the chapter. I affirm that Angela was lead author and fully involved in setting up and conducting the research, obtaining the data and analyzing results as well as preparing and writing the material presented in the article.

Sincerely,

[Signature]

Emeritus Professor of
Applied Linguistics & Discourse Studies

Dated

Appendix B: Statement from supervisors

July 21, 2023

To Whom It May Concern:

As the co-supervisors of Angela Carolina de Moraes Garcia's doctoral thesis, we attest that she is the author of this dissertation and was the lead author in all the published manuscripts included herein. She was fully involved in setting up and conducting the research, obtaining data and analyzing results, as well as preparing and writing the material presented in both the solely authored and co-authored chapter/articles integrated in the thesis.

Further, I attest that I have reviewed and confirmed information presented by Angela de Moraes Garcias in the Preface with regard to her specific contribution and those of all other collaborators or co-authors of published manuscripts integrated in this dissertation.

Sincerely,

Guillaume Gentil, Ph.D.
Co-supervisor/Professor

Angel Arias
Co-supervisor/Assistant Professor
Appendix C: Supervisor Review and Approval Form (Chapter 5)

For consideration of submitted ethics protocols, the Carleton University Research Ethics Boards require evidence that all student protocol documents (i.e., undergraduate, graduate and post-doctoral fellows) have been reviewed and approved by a faculty supervisor or sponsor.

Instructions:
After filling out the details in the text below, faculty sponsor/supervisors should either
1) Print and scan this document, or
2) Email the text of the document (below; with signature optional) to the lead researcher

The form or email can then be uploaded (in PDF format) with the protocol, to the CUResearch.

As the faculty supervisor or sponsor, I confirm that I have reviewed and that I approve for submission for ethics review, the protocol entitled Name of Protocol from Name of Lead Researcher on Date

Signature of Faculty Supervisor/Sponsor: ____________________________

Name of Faculty Supervisor/Sponsor: Natasha Artemova

Date: 10 October 2019
Certificate of Completion

This document certifies that

Angela Carolina de Moraes Garcia

has completed the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans Course on Research Ethics (TCPS 2: CORE)

Date of Issue: 14 October, 2019
Appendix E: Email invitation (Chapter 5, Phase 1)

Subject: Invitation to participate in a research project on Assessment of pilots’ listening ability

Hello,

My name is Angela Garcia, and I am a PhD student in the School of Linguistics and Language Studies at Carleton University. I am working on a research project under the supervision of Professors Janna Fox and Natasha Artemeva.

I am writing to you today to invite you to participate in Phase 1 of a study entitled “Contexts and constructs: Implications for the testing of listening in pilots’ radiotelephony communication with the air traffic controllers”. This study seeks to answer the following research question: “What are the characteristics of the listening tasks that pilots undertake during radiotelephony communications?”

This study aims at contributing to the improvement of tests that address the Civil Aviation Organization (ICAO)’s language proficiency requirements for pilots and air traffic controllers, especially in relation to the assessment of listening. This study is divided into two phases. In phase 1, participants are asked to respond to a questionnaire. The participants in Phase 1 must be airplane or helicopter pilots. At the end of the questionnaire, participants who have experience working as an ICAO test rater will be advised to contact the researcher via email (angelagarcia@cmail.carleton.ca) in case they wish to participate in Phase 2 of the study. In Phase 2, participants will be interviewed. The participants in Phase 2 may be a pilot or an air traffic controller who works or has worked as an ICAO language proficiency test rater, an ICAO language proficiency test developer, or a language assessment expert who has experience with aviation testing.

Phase 1 of this study involves one 20-minute questionnaire that will take place online, through the Qualtrics Survey Software, an online survey platform.

Care will be taken to protect your identity. This will be done by keeping all responses anonymous.

All research data will be encrypted. Research data will only be accessible by the researcher and the research supervisor.

The ethics protocol for this project was reviewed by the Carleton University Research Ethics Board, which provided clearance to carry out the research. (Clearance expires on: November 30, 2020.)

If you have any ethical concerns with the study, please contact the Carleton University Research Ethics Board-A (by phone at 613-520-2600 ext. 2517 or via email at ethics@carleton.ca).

If you would like to participate in this research project, please click on the link below.

https://carletonu.az1.qualtrics.com/jfe/form/SV_6hBQFzJzSRrSj4N

If you have any questions, please contact me at:

angelagarcia@cmail.carleton.ca

Sincerely,

Angela Carolina de Moraes Garcia.
Appendix F: Online advertisement (Chapter 5, Phase 1)
(Used on LinkedIn, Facebook, Instagram, and WhatsApp)

Participate in a study on pilots’ listening assessment!
To participate in this study, you must be:

✓ An airplane or helicopter licensed pilot
✓ Comfortable in the English language

This study is composed of two phases. This invitation is to take part in Phase 1, which consists of a 20-minute questionnaire. You will be asked questions about the listening tasks performed by pilots in their communications with the air traffic controllers on the radiotelephone.

Participation is voluntary. Data will be anonymous.

To participate in this study, please click here:
https://carletonu.az1.qualtrics.com/jfe/form/SV_6hBQFzJzSRrSj4N

For more information about this study, please contact the researcher at angelagarcia@cmail.carleton.ca.

The ethics protocol for this project has been reviewed and cleared by the Carleton University Research Ethics Board. If you have any ethical concerns with the study, please contact the Carleton University Research Ethics Board-A (by phone at 613-520-2600 ext. 2517 or via email at ethics@carleton.ca).

Thank you very much!
Subject: Invitation to participate in a research project on Assessment of pilots’ listening ability

Hello,

My name is Angela Garcia, and I am a PhD student in the School of Linguistics and Language Studies at Carleton University. I am working on a research project under the supervision of Professors Janna Fox and Natasha Artemeva.

I am writing to you today to invite you to participate in Phase 2 of a study entitled “Contexts and constructs: Implications for the testing of listening in pilots’ radiotelephony communication with the air traffic controllers”. This study seeks to answer the following research question: “What are the characteristics of the listening tasks that pilots undertake during radiotelephony communications?”

This study aims at contributing to the improvement of tests that address the Civil Aviation Organization (ICAO)’s language proficiency requirements for pilots and air traffic controllers, especially in relation to the assessment of listening. This study is divided into two phases. In phase 1, participants were asked to respond to a questionnaire. The participants in Phase 1 consisted of airplane or helicopter pilots. At the end of the questionnaire, participants who have experience working as an ICAO test rater were asked to contact the researcher via email (angelagarcia@cmail.carleton.ca) in case they wished to participate in Phase 2 of the study. Participation in phase 2 is also being advertised on social media. In Phase 2, participants will be interviewed. The participants in Phase 2 may be a pilot or an air traffic controller who works or has worked as an ICAO language proficiency test rater, an ICAO language proficiency test developer, or a language assessment expert who has experience with aviation testing.

Phase 2 of this study involves one 30-minute interview that will take place online (through Skype). With your consent, the interviews will be audio-recorded. Care will be taken to protect your identity. You will be assigned a code/pseudonym so that the data are fully anonymized. Once the recording has been transcribed, the audio-recording will be destroyed. Participants are allowed to request certain responses not to be included in the final project.

You will have the right to end your participation in the study at any time, for any reason, within seven days after the interview has been conducted. If you choose to withdraw, all the information you have provided will be destroyed.

All research data, including audio-recordings and any notes will be encrypted. Research data will only be accessible by the researcher and the research supervisor.

The ethics protocol for this project was reviewed by the Carleton University Research Ethics Board, which provided clearance to carry out the research. (Clearance expires on: November 30, 2020.)

If you have any ethical concerns with the study, please contact the Carleton University Research Ethics
Board-A (by phone at 613-520-2600 ext. 2517 or via email at ethics@carleton.ca).

If you would like to participate in this research project, or have any questions, please contact me at angelagarcia@email.carleton.ca

Sincerely,

Angela Carolina de Moraes Garcia.
Appendix H: Online Advertisement (Chapter 5, Phase 2)

Participate in a study on pilots’ listening assessment!

To participate in this study, you must be:

✓ An airplane or helicopter pilot who has experience working as an ICAO test rater; or
✓ An air traffic controller who has experience working as an ICAO test rater; or
✓ An ICAO test developer; or
✓ A language testing expert who has experience with aviation English testing (Masters in Language Testing or PhD in Applied Linguistics with specialization in Language Testing); and
✓ You must be fluent in English.

This study is composed of two phases. This invitation is to take part in Phase 2, which consists of a 30-minute questionnaire. You will be asked questions about the listening tasks performed by pilots in their communications with the air traffic controllers on the radiotelephone.

Participation is voluntary. Data will be anonymized.

To participate in this study, or if you want more information about it, please contact the researcher at angelagarcia@cmail.carleton.ca
Thank you very much!
Appendix I: Informed Consent Form (Chapter 5, Phase 1)

Name and Contact Information of Researchers

Angela Carolina de Moraes Garcia, Carleton University, School of Linguistics and Language Studies/Faculty of Arts and Social Sciences
Tel.: 613-520-6612
Email: angelagarcia@cmail.carleton.ca

Supervisors and Contact Information:
Dr. Janna Fox
Tel: 613-520-2600
Email: janna_fox@carleton.ca
Dr. Natasha Artemeva
Tel: 613-520-2600
Email: natasha.artemeva@carleton.ca

Project Title

Contexts and constructs: Implications for the testing of listening in pilots’ radiotelephony communication with the air traffic controllers

Project Sponsor and Funder (if any)

Not applicable

Carleton University Project Clearance
Clearance #: 111875 Date of Clearance: November 19, 2019

Invitation

You are invited to take part in this research project because you are an airplane/helicopter commercial or airline transport pilot. The information in this form is intended to help you understand what we are asking of you so that you can decide whether you agree to participate in this study. Your participation in this study is voluntary, and a decision not to participate will not be used against you in any way. As you read this form, and decide whether to participate, please ask all the questions you might have, take whatever time you need, and consult with others as you wish.

This study is divided into two phases (Phase 1 and Phase 2). This informed consent relates to Phase 1 only.
What is the purpose of the study?
This study aims to contribute to the understanding of the nature of the listening tasks that pilots perform when communicating with air traffic controllers on the radiotelephone. This will help International Civil Aviation Organization (ICAO)’s language proficiency test developers to design listening tests and questions that are more appropriate to the language that the pilots use in their real world, and it will contribute to more valid testing practices.

What will I be asked to do?
If you agree to take part in Phase 1 of this study, we will ask you to complete an online questionnaire. Completing the questionnaire will take you approximately 20 minutes. If you have experience working as an ICAO test rater, at the end of the questionnaire, you will be asked whether you wish to participate in Phase 2, which will consist of voice-only Skype interviews.

Risks and Inconveniences
The probability of mild psychological and social/economic risk is very low. Remember that your participation is voluntary and that the data are anonymous. You can exit the questionnaire any time before submitting it.

Possible Benefits
You may not receive any direct benefit from your participation in this study. However, your participation may allow researchers to better understand the nature of listening in the context of pilot/air traffic controller communications, as well as suggest to language test developers what would be the most appropriate listening tasks to be included in an aviation language test for pilots. We believe that the results of this research will contribute to more valid testing.

Compensation/Incentives
Your participation is voluntary. You will not be paid or compensated for your participation in this study.

No waiver of your rights
By signing this form, you are not waiving any rights or releasing the researchers from any liability.

Withdrawing from the study
You can exit the questionnaire at any time before submitting it. In this case, no information will be recorded. After you submit the questionnaire, it is not possible to remove your answers from the study data.

Confidentiality
We will treat your personal information as confidential. Questionnaire data are anonymous. You can exit the questionnaire at any time before submitting it, and no information will be recorded. The researcher will not collect any identifying information, including IP addresses. Research records may be accessed by the Carleton University Research Ethics Board in order to ensure continuing ethics compliance.

The results of this study may be published or presented at an academic conference or meeting, but the data will be presented so that it will not be possible to identify any participants.
Data Retention

Your data will be stored and protected by Qualtrics on servers in Toronto, Canada.

New information during the study

In the event that any changes could affect your decision to continue participating in this study, you will be promptly informed.

Ethics review

This project was reviewed and cleared by the Carleton University Research Ethics Board - A. If you have any ethical concerns with the study, please contact Carleton University Research Ethics Board (by phone at 613-520-2600 [ext. 2517] or by email at ethics@carleton.ca).

Statement of consent

I voluntarily agree to participate in this study. □ Yes □ No
Appendix J: Informed Consent Form (Chapter 5, Phase 2)

Name and Contact Information of Researchers:

Angela Carolina de Moraes Garcia, Carleton University, School of Linguistics and Language Studies/Faculty of Arts and Social Sciences

Tel.: 613-520-6612

Email: angelagarcia@cmail.carleton.ca

Supervisor and Contact Information:

Dr. Janna Fox

Tel: 613-520-2600

Email: janna_fox@carletonv.ca

Dr. Natasha Artemeva

Tel: 613-520-2600

Email: natasha.artemeva@carleton.ca

Project Title

Contexts and constructs: Implications for the testing of listening in pilots’ radiotelephony communication with the air traffic controllers

Project Sponsor and Funder (if any)

Not applicable.

Carleton University Project Clearance

Clearance #: 111875 Date of Clearance: November 19, 2019
Invitation

You are invited to take part in a research project because you are either a pilot or an air traffic controller who has experience as an International Civil Aviation Organization (ICAO) language proficiency test rater, an ICAO language proficiency test developer, or language testing expert who has experience with aviation English testing. The information in this form is intended to help you understand what we are asking of you so that you can decide whether you agree to participate in this study. Your participation in this study is voluntary, and a decision not to participate will not be used against you in any way. As you read this form, and decide whether to participate, please ask all the questions you might have, take whatever time you need, and consult with others as you wish.

This study is divided into two phases (Phase 1 and Phase 2). This informed consent relates to Phase 2 only.

What is the purpose of the study?

This study aims to contribute to the understanding of the nature of the listening tasks that pilots perform when communicating with air traffic controllers on the radiotelephone. This will help ICAO’s language proficiency test developers to design listening tests and questions that are more appropriate to the language that the pilots use in their real world, and it will contribute to more valid testing practices.

What will I be asked to do?

If you agree to take part in Phase 2 of this study, we will first ask you to print, sign and send to the researcher (angelagarcia@cmail.carleton.ca) this Consent Form. We will then schedule a Skype one-on-one voice-only interview with you. The interview will be audio recorded, but you will not be asked to say your name. You will be assigned an alphabetical code instead (for example, Participant A, Participant B, etc.). In the interview, you will be asked questions related to your background, experience, knowledge, thoughts and opinions about the listening tasks that pilots usually perform in radiotelephony communications, and also questions related to the assessment of pilots’ listening abilities. The interview will take approximately 30 minutes.
**Why will the interview be audio recorded?**

Audio recording is a condition of participation in the study. The interview needs to be audio recorded so that the researcher can transcribe it later and analyze the data. As explained, you will be assigned a code (for example, “Participant A”). Your name will not be recorded. After the interview is transcribed and the transcripts verified, the recording will be destroyed. Audio recording is not optional. If you do not consent to be audio recorded, it will not be possible to participate in this phase of the study. Remember that your participation is voluntary and that the data will be anonymized.

**Risks and Inconveniences**

The probability of mild psychological and social/economic risk is very low. Remember that your participation is voluntary and that the data are anonymous. You can ask the researcher to withdraw from the study within 7 (seven) days after the interview. In this case, their recordings, and transcriptions (if ready or partially ready) will be destroyed.

**Possible Benefits**

You may not receive any direct benefit from your participation in this study. However, your participation may allow researchers to better understand the nature of listening in the context of pilot/air traffic controller communications, as well as to suggest to language test developers what would be the most appropriate listening tasks to be included in an aviation language test for pilots. We believe that the results of this research will contribute to more valid testing.

**Compensation/Incentives**

Your participation is voluntary. You will not be paid or compensated for your participation in this study.

**No waiver of your rights**

By signing this form, you are not waiving any rights or releasing the researchers from any liability.

**Withdrawing from the study**
You can ask the Principal Investigator (named above) for withdrawal within 7 (seven) days after your interview. All the data collected from you will be destroyed upon your withdrawal. After that time, you will not be able to withdraw.

**Confidentiality**

As mentioned, you will be assigned a code so that your identity will not be directly associated with the data you have provided. After the interview is transcribed, the recording will be destroyed. Your data will be anonymized. No information that discloses your identity will be released or published. Research records may be accessed by the Carleton University Research Ethics Board in order to ensure continuing ethics compliance.

The results of this study may be published or presented at an academic conference or meeting, but the data will be presented so that it will not be possible to identify any participants.

All data, including coded information, will be kept in a password-protected file on a password-protected secure computer. The researcher will password protect any research data that will be stored or transferred.

**Data Retention**

After the study is completed, your de-identified data will be retained for future research use for five years. Then, all data will be destroyed.

**New information during the study**

In the event that any changes could affect your decision to continue participating in this study, you will be promptly informed.

**Ethics review**

This project was reviewed and cleared by the Carleton University Research Ethics Board A. If you have any ethical concerns with the study, please contact Carleton University Research Ethics Board (by phone at 613-520-2600 [ext. 2517], or by email at ethics@carleton.ca).

**Statement of consent – print, sign and send to the researcher**

(angelagarcia@cmail.carleton.ca)

I voluntarily agree to participate in this study. ___Yes ___No
I agree to be audio recorded (not optional)  ___Yes    ___No

________________________  ________________________
Signature of participant      Date
Appendix K: Interview Guide (Chapter 5, Phase 2)

Interviewer: Good morning/good afternoon/good evening. It’s Angela Garcia. How are you?

Participant: …

Interviewer: First of all, I would like to thank you very much for participating in this study. I really appreciate it. I would like to remind you that you are being recorded, and that the recording will be transcribed and then destroyed. I have assigned you a code as Participant ___ (A to F). If you are not sure about an answer, you don’t need to respond. Just say you are not sure about it. Do you have any questions before we start?

Participant: …

Interviewer: What is your first language?

Participant: …

Interviewer: (if participant is a pilot) What pilot’s license do you hold? How many flight hours do you have? Approximately how many hours have you flown internationally? (if non-native speaker) Have you ever had your language proficiency assessed according to the ICAO Language Proficiency Requirements? Do you mind telling me the level you were awarded? How long have you been working as a rater?

Participant: …

Interviewer: (if participant is an air traffic controller) What license do you hold? How much experience do you have working as an air traffic controller? How much experience do you have controlling international flights? (if non-native speaker) Have you ever had your language proficiency assessed according to the ICAO Language Proficiency Requirements? Do you mind telling me the level you were awarded? How long have you been working as a rater?

Participant: …

Interviewer: Tell me a little bit (more) about your experience with aviation English.

Participant: …

Interviewer: So, what do you think should be assessed in terms of pilot’s listening comprehension in the radiotelephony communications?

Participant: …

Interviewer: The ICAO language proficiency requirements require only plain English to be assessed. Do you think comprehension of standardized phraseology should also be assessed? Why (not)?

Participant: …

Interviewer: Do you think comprehension of plain English should also be assessed? Why (not)?

Participant: …

Interviewer: What listening abilities do you consider to be important to be assessed (for example, understanding main ideas, recognizing words and numbers, understanding different accents, recognizing indirect meaning like politeness, sarcasm, professionalism, anger, etc., extracting specific meaning
from both short and long transmissions, understanding speech at different speed, understanding the meaning of uncommon words and expressions)? Why (not)?

Participant: …

Interviewer: Do you think it is common for pilots to struggle to understand ATCOs’ pronunciation? Why/how?

Participant: …

Interviewer: Is it important for pilots to understand more complex grammar in order to understand an air traffic controller communication?

Participant: …

Interviewer: Is it important for pilots to be able to understand unfamiliar vocabulary? Why (not)? Can you give me some examples?

Participant: …

Interviewer: What about idioms and phrasal verbs?

Participant: …

Interviewer: Do you think ATCOs speak faster than the recommended 100 words per minute? If necessary: how much faster?

Participant: …

Interviewer: Do pilots interact more with non-native speakers of English than with native speakers when conducting international flights? Why (not)?

Participant: …

Interviewer: To what extent is it important to assess a pilot’s ability to understand different accents? Why?

Participant: …

Interviewer: To what extent is it important to assess a pilot’s ability to understand cultural subtleties? How can the ability to understand cultural subtleties be assessed?

Participant: …

Interviewer: Speaking of the real world, what is the maximum number of pieces of information that an air traffic controller communication might include in a single transmission?

Participant: …

Interviewer: In your opinion, what are the main challenges that pilots face when trying to understand an air traffic controller communication?

Participant: …

Interviewer: How do you think pilots’ listening ability should be assessed? Should it be assessed separately from the speaking test, together with the speaking test or both?
Participant: …

Interviewer: (if a rater) As a rater, what is your opinion about the descriptors for comprehension in the ICAO rating scale? Is there something you think that should be changed, deleted or added?

Participant: …

Interviewer: Do you think that comprehension should have its own set of scales, with descriptors for different categories? If so, what categories should it include?

Participant: …

Interviewer: Is there anything that I didn’t ask that you think I should have asked?

Participant: …

Interviewer: That’s it, thank you very, very much for your participation. Do you want to receive a summary of the results of this study?

Participant: …

Interviewer: (If candidate says “yes”) ok. As soon as the results are analyzed, I will send you an email with the summary.

Thank you very much once again! Have a wonderful day/afternoon/evening/week. Bye!
Appendix L: Ethics Clearance Certificate (Chapter 5)

CERTIFICATION OF INSTITUTIONAL ETHICS CLEARANCE

The Carleton University Research Ethics Board-A (CUREB-A) has granted ethics clearance for the research project described below and research may now proceed. CUREB-A is constituted and operates in compliance with the *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans* (TCPS2).

Ethics Protocol Clearance ID: Project # 111875

Project Team Members: Ms. Angela Garcia (Primary Investigator)
Dr. Natalia Artemeva (Research Supervisor)
Prof. Janna Fox (Research Supervisor)

**Project Title:** Contexts and constructs: Implications for the testing of listening in pilots’ radiotelephony communication with the air traffic controllers

**Funding Source** (If applicable):


Please ensure the study clearance number is prominently placed in all recruitment and consent materials: CUREB-A Clearance # 111875.

**Restrictions:**

This certification is subject to the following conditions:

1. Clearance is granted only for the research and purposes described in the application. Any modification to the approved research must be submitted to CUREB-A via a Change to Protocol Form. All changes must be cleared prior to the continuance of the research.

2. An Annual Status Report for the renewal of ethics clearance must be submitted and cleared by the renewal date listed above. Failure to submit the Annual Status Report will result in the closure of the file. If funding is associated, funds will be frozen.

3. A closure request must be sent to CUREB-A when the research is complete or terminated.
5. During the course of the study, if you encounter an adverse event, material incidental finding, protocol deviation or other unanticipated problem, you must complete and submit a Report of Adverse Events and Unanticipated Problems Form, found here: https://carleton.ca/researchethics/forms-and-templates/

Failure to conduct the research in accordance with the principles of the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans 2nd edition and the Carleton University Policies and Procedures for the Ethical Conduct of Research may result in the suspension or termination of the research project.

Upon reasonable request, it is the policy of CUREB, for cleared protocols, to release the name of the PI, the title of the project, and the date of clearance and any renewal(s).

Please contact the Research Compliance Coordinators, at ethics@carleton.ca, if you have any questions.

CLEARED BY:

Date: November 19, 2019

Janet Mantler, PhD, Vice-Chair, CUREB-A