LANGUAGE SWITCHING IN AVIATION

A thesis presented in partial fulfilment of the requirements for the degree of
Doctor of Philosophy in Aviation
at Massey University, Manawatū, New Zealand.

Martina Daskova

2018

Eternal rest grant unto the victims of aircraft accidents, O Lord, and let perpetual light shine upon them. For the sake of Your sorrowful passion, may their souls rest in peace. Amen

Abstract

Clear and precise communication between pilots and air traffic controllers is a precondition for safe operations. Communication has long been identified as a major element of the cockpit–controller interface, explaining one third of general aviation incidents (Etem & Patten, 1998). Yet, despite multilingualism with English as the *lingua franca* being a characteristic of aviation communication, little research appears to have investigated the efficiency of operation of bilinguals alternating between their dominant, usually native, language and English in a bilingual air traffic environment.

The studies undertaken for this research sought to rectify this situation by examining the cognitive aspects of situation awareness during language switching in aviation. Quantitatively and qualitatively analysed responses to an online-distributed survey aimed at investigating the current bilingual situation in aviation revealed that while situation awareness for the majority (76%) of native-English speakers was adversely affected by bilingualism, almost 30% of bilinguals also reported their situation awareness being affected. Subsequent experimental analyses using a language switching paradigm investigated how participants recognize a target call sign, identify an error and predict in bilingual compared with monolingual English conditions. The effect of the language condition participants' native Chinese only, English only, or a mix of both, varied across the three tasks. Call sign recognition performance was found to be faster in the English condition than in the bilingual condition, but accuracy did not differ, a finding that was attributed to the effect of call sign similarity. However, when the task was more complicated, the difference between the conditions diminished. No effect on performance was found for simultaneously listening to two speech sources, which is potentially analogous to cockpit communication and radio calls. The error analyses served to test for response bias by calculating sensitivity, *d'*, and decision criterion *C* in accordance with Stanislaw and Todorov's (1999) Signal Detection Theory calculations.

Several cognitive implications for practice were proposed, for example, in Crew Resource Management (CRM) training and personal airmanship development, exploration of own behavioural biases might be used to adjust the placement of the criterion. The cognitive implications largely focused on affecting attitudes to increase awareness. Attention was focused on performance of bilinguals to identify which language condition facilitated faster and more accurate responses. The findings were unable to support any of the conditions, leaving the question: *Would a universal language for communication on radio frequencies be worth considering, to allow everyone to understand what is said?* Disentangling the effects of language switching on the performance of bilingual pilots and air traffic controllers remains a task for future studies.

Acknowledgements

I would like to express my appreciation to several people for their support and advice over the last three years on my PhD journey. I hold a deep sense of gratitude to my main supervisor, Dr Andrew Gilbey, for his patience, who, despite my shortcomings, helped me to persevere on the right research path. My sincere thanks belongs also to Dr Michael MacAskill from the New Zealand Brain Research Institute, who helped me with the development of the experiments using the PsychoPy software. I am grateful to my second supervisor, Dr Savern Rewetti, and to the entire School of Aviation, for supporting me in all my milestones; whether it was time to recruit participants or was there an opportunity to share my research findings at a conference. I would also like to thank Sarah Fifield for her valuable and insightful comments and suggestions to improve the quality of the thesis. Special thanks to Sherryn Irvine for her support, not only in a work environment, but also for sharing with me a loving atmosphere within her community of friends. Last, but not least, I am grateful to my entire family, who always pooled their resources so that I would be able to complete my studies in comfort.

I give thanks to my best friend, Lord Jesus Christ. *Terima kasih*, Lord, my Hero, for the gift of all the people during these years and for everything You do for us, within us, and through us. Pardon me as well, for too often I am not even aware of how much you help me, how much You unceasingly care. Mistakes are mine, the good comes from the Lord Jesus Christ.

I wanted to fly; You removed the solid ground.

CONTENTS

Chapter One 1

INTRODUCTION

1.1. Background of the Study 1

1.2. Thesis Outline..4

Chapter Two..6

THE STUDY CONTEXT

2.1. Bilingualism: Definition and Use 6

2.2. Bilingual Air Traffic Environment 7

2.3. Bilingualism as a Contributing Factor in Safety Occurrences..............................8

2.4. Bilingual Air Traffic Conflict 10

2.5. English Language Proficiency Regulation in Aviation 12

2.6. Language Issues in Aviation after ICAO LPR Implementation..........................13

Chapter Three..17

LITERATURE REVIEW

Part I: Defining the Concepts

3.1. Overview 17

3.2. The Importance of SA in Aviation 17

3.3. Bilingualism, SA and Cognition...23

3.4. Bilingual IFR Communications Simulation Studies 27

3.5. Language Switching 29

3.6. Contrasting Chinese and English Language 34

3.6.1. Pronunciation and Accent...36

3.6.2. Number of Instructions in One Transmission...37

3.6.3. Grammar Issues 37

3.7. Numerical Processing...39

Part II: Methodological Considerations

3.8. Methodological Considerations..44

3.9. The Reason for Including Signal Detection Theory..46

3.10. SDT, SA, and Bilingualism 47

3.11. Yes–No Procedure 49
3.12. Response Time: The Challenge 50
3.13. Markers for Language Control 53
3.14. Discriminability Index and Decision Criterion 55
3.15. Background Noise and Speech to Noise Ratio 59

Part III: Summary

3.16. Summary of the Literature Review 60
3.17. Identification of the Research Problem 62
3.18. Identification of the Research Questions 62

Chapter Four 64

STUDY 1: Pilot and ATCO Current Language Experiences

4.1. Introduction 64
4.2. Method 66
4.2.1. Participants 66
4.2.2. Materials 66
4.2.3. Procedure 71
4.3. Results 72
4.3.1. General findings 72
4.3.2. Bilingual Air Traffic Environment 74
4.3.3. Perceived Effects on SA 76
4.3.4. Experienced Consequences 78
4.4. Discussion 81

Chapter Five 87

STUDY 2: Call Sign Recognition

5.1. Introduction 87
5.1.1. Call Sign Aviation Regulation and Research 87
5.2. Method 89
5.2.1. Overview 89
5.2.2. Participants 89
5.2.3. Design 90
5.2.4. Materials 93
5.2.5. Procedure 99

5.3. Results 100
5.3.1. General Findings 100
5.3.2. Performance Speed 104
5.3.3. Performance Accuracy 110
5.3.4. SDT Measures 113
5.4. Discussion 114

Chapter Six 120
STUDY 3: Error Identification
6.1. Introduction 120
6.1.1. Detection of Erroneous Message 121
6.2. Method 124
6.2.1. Overview 124
6.2.2. Participants 124
6.2.3. Design 125
6.2.4. Materials 127
6.2.5. Procedure 131
6.3. Results 132
6.3.1. General findings 132
6.3.2. Performance Speed 133
6.3.3. Performance Accuracy 141
6.3.4. SDT Measures 143
6.4. Discussion 144

Chapter Seven 149
STUDY 4: Prediction
7.1. Introduction 149
7.1.1. Many Terms, Same Concept? Anticipation, Prediction, Expectation, Projection 150
7.1.2. Number Series 152
7.2. Method 153
7.2.1. Overview 153
7.2.2. Participants 153
7.2.3. Design 154
7.2.4. Materials 157

7.2.5. Procedure 161
7.3. Results 162
7.3.1. General Findings 162
7.3.2. Performance Speed 165
7.3.3. Performance Accuracy 169
7.3.4. SDT Measures 171
7.4. Discussion 172

Chapter Eight 176
STUDY 5: Listening To Radio Calls Over Background Talk
8.1. Introduction 176
8.2. Method 180
8.2.1. Overview 180
8.2.2. Participants 181
8.2.3. Design 181
8.2.4. Materials 186
8.2.5. Procedure 191
8.3. Results 192
8.3.1. General Findings 192
8.3.2. Performance Speed 194
8.3.3. Performance Accuracy 197
8.3.4. SDT Measures 201
8.4. Discussion 202

Chapter Nine 209
STUDY 6: Sterile Cockpit
9.1. Introduction 209
9.2. Method 211
9.2.1. Overview 211
9.2.2. Considerations for the Analysis 211
9.2.3. Measures 212
9.3. Results 213
9.3.1. Overview 213
9.3.2. Performance Speed 213

9.3.3. Performance Accuracy 216
9.3.4. SDT Measures 217
9.4. Discussion 218

Chapter Ten 223

GENERAL DISCUSSION

10.1. Summary 223
10.2. Discussion 224
10.3. Limitations and Future Research 229
10.4. Cognitive Implications for Practice 232

CONCLUSION 236

REFERENCES 238

Appendix A: Low-Risk Ethics Notification for Study 1 270
Appendix B: Low-Risk Ethics Notification for Study 2 271
Appendix C: Stimuli List for Study 2 272
Appendix D: Low-Risk Ethics Notification for Study 3 273
Appendix E: Stimuli List for Study 3 274
Appendix F: Low-Risk Ethics Notification for Study 4 275
Appendix G: Stimuli List for Study 4 276
Appendix H: Low-Risk Ethics Notification for Study 5 277
Appendix I: Stimuli List for Study 5 278
Appendix J: Transcript of Background Talks for Study 5 279

List of Tables

Table 1 *Information–Response Matrix* 49

Table 2 *Demographic: First Language of Participants by Language Families, and Percentage of Participants Speaking more than Two Languages across Corresponding Language Families* 74

Table 3 *Survey Participants' Experiences of Countries Using Bilingual* 75

Table 4 *Perceived Effects of Language Alternation on Situation Awareness of Native- and Non-Native English-Speaking Participants* 76

Table 5 *Reported Frequency of Experienced Struggles to Speak English by Non-Native English-Speaking Participants* 78

Table 6 *The Consequences of Difficulties Related to Bilingual Air Traffic* 79

Table 7 *ICAO Rating Scale, CEFR, and IELTS Comparison Chart* 98

Table 8 *Mean Response Times (RT), Mean Durations of Stimuli, and Mean Pure RT in Seconds in First (L1), Second (L2), and Language Switching (Mix) Conditions* 103

Table 9 *Median (Mdn) Pure Response Times in Seconds across Language Condition and Inter-stimuli Interval (ISI) Factors* 107

Table 10 *Median (Mdn) Pure Response Times in Seconds across Language Condition and Similarity Factors* 107

Table 11 *Median (Mdn) Pure Response Times in Seconds on Chinese and English Word Stimuli in Monolingual (L1 and L2) and Language Switching (Mix) Conditions: Mixing Costs and Switch Costs* 108

Table 12 *Mean Pure Response Times in Seconds across the English Language Proficiency Levels* 109

Table 13 *Error Types (Miss and False Alarm), Hits and Correct Rejections (CR), and Total Number of Errors across Language Conditions, and Percentage of Errors from 5100 Stimuli (%Error$_T$)* 110

Table 14 *Number (n$_{errors}$) and Percentage of Errors (%Error$_E$) across ISI Levels and as a Proportion of Total Number of Stimuli (n$_{stimuli}$) in Each ISI Level (%Error$_{ISI}$)* 111

Table 15 *Number (n$_{errors}$) and Percentage of Errors (%Error$_E$) across Levels of Similarity Factor (%Error$_S$), and as a Proportion of Total Number of Stimuli (n$_{stimuli}$) in Each Level of Similarity Factor (n$_{stimuli}$)* 112

Table 16 *Frequency of Correct Responses and Errors, and Percentage of Errors (%Errors) across English Language Proficiency Levels* 113

Table 17 *Sensitivity (d') and Decision Criterion (C) of Call Sign Recognition Task across the Three Language Conditions: Native (Chinese) Language (L1), Second (English) Language (L2), and Language Switching (Mix)* 113

Table 18 *Mean Response Times (RT), Mean Durations of Stimuli, and Mean Pure RT in Seconds, in First (L1), Second (L2), and Language Switching (Mix) Conditions* 133

Table 19 *Median (Mdn) Pure Response Times in Seconds across Language Condition and Inter-stimuli Interval (ISI) Factors* 136

Table 20 *Median (Mdn) Pure Response Times in Seconds on Correct Rejections (CR) and Hits across the Language Condition Factor* 137

Table 21 *Median (Mdn) Pure Response Times in Seconds on Chinese and English Word Stimuli for Correct Rejections (CR), Hits and Both Combined in Monolingual (L1 and L2) and Language Switching (Mix) Conditions; Mixing Costs and Switch Costs*. 138

Table 22 *Mean Pure Response Times in Seconds across the English Language Proficiency Levels* 139

Table 23 *Error Types (Miss and False Alarm), Hits and Correct Rejections (CR), Total Number of Errors and "Say Again" Requests across Language Conditions, and Percentage of Errors from 5200 stimuli (%Error$_T$)* 141

Table 24 *Number (n_{errors}) and Percentage of Errors (%Error$_E$) across ISI Levels and as a Proportion of Total Number of Stimuli ($n_{stimuli}$) in Each ISI Level (%Error$_{ISI}$)* 142

Table 25 *Number of Participants (n), Correct Responses and Errors, and Percentage of Errors (%Errors) across English Language Proficiency Levels* 143

Table 26 *Sensitivity (d') and Decision Criterion (C) of Error Identification Task across the Three Language Conditions: Native (Chinese) Language (L1), Second (English) Language (L2), and Language Switching (Mix)* 144

Table 27 *Mean Response Times (RT), Mean Durations of Stimuli, and Mean Pure RT in Seconds in First (L1), Second (L2), and Language Switching (Mix) Conditions* 164

Table 28 *Mean Pure Response Times in Seconds and Standard Deviations (SD) across Language Condition and Position Factors* 167

Table 29 *Mean Pure Response Times in Seconds and Standard Deviations (SD) across Language Condition and Pattern Factors* 167

Table 30 *Mean Pure Response Times in Seconds across the English Language Proficiency Levels* 168

Table 31 *Error Types (Miss and False Alarm), Hits and Correct Rejections (CR), and Total Number of Errors across Language Conditions, and Percentage of Errors from 1920 Stimuli (%Error$_T$)* 169

Table 32 *Number (n_{errors}) and Percentage of Errors (%Error$_E$) across Levels of Position Factor and as a Proportion of Total Number of Stimuli in Each Position Factor Level (%Error$_P$)* 170

Table 33 *Number (n_{errors}) and Percentage of Errors (%Error$_E$) across Pattern Factor Levels and as a Proportion of Total Number of Stimuli in Each Pattern Factor Level (%Error$_P$)* 170

Table 34 *Number of Participants (n), Correct Responses and Errors, and Percentage of Errors (%Errors) across English Language Proficiency Levels* 171

Table 35 *Sensitivity (d') and Decision Criterion (C) of Prediction Task across the Three Language Conditions: Native (Chinese) Language (L1), Second (English) Language (L2), and Language Switching (Mix)* 171

Table 36 *Mean Response Times (RT), Mean Durations of Stimuli, and Mean Pure RT in Seconds in Second (L2) and Language Switching (Mix) Conditions across the Three Tasks with Background Talk (Call Sign Recognition, Error Identification, and Prediction)* 193

Table 37 *Mean Pure Response Times in Seconds across Language Condition and Background Talk Factors* 194

Table 38 *Median (Mdn) Pure Response Times in Seconds on Chinese and English Word Stimuli in Monolingual (L2) and Language Switching (Mix) Conditions; Mixing Costs and Switch Costs* 195

Table 39 *Mean Pure Response Times in Seconds and Standard Deviations (SD) across Language Condition and Background Talk Factors* 196

Table 40 *Median (Mdn) Pure Response Times in Seconds on Chinese and English Word Stimuli in Monolingual (L2) and Language Switching (Mix) Conditions; Mixing Costs and Switch Costs* 196

Table 41 *Mean Pure Response Times in Seconds and Standard Deviations (SD) across Language Condition and Background Talk Factors* 197

Table 42 *Number and Percentage of Errors across the Tasks and Background Talks* 198

Table 43 *Number and Percentage of Errors across the Tasks and Language Conditions, Total Number of Errors (Totals), Total Number of Stimuli within Tasks (N), and Percentage of Total Number of Errors (%ErrorT) within Tasks*... 199

Table 44 *Distribution of Errors (Miss and False Alarm), Hits and Correct Rejections (CR) in Call Sign Recognition Task, Total Number of Errors across Language Conditions, and Percentage of Errors from 1350 Stimuli (%Error$_T$)*... 200

Table 45 *Distribution of Errors (Miss and False Alarm), Hits and Correct Rejections (CR) in Error Identification Task, Total Number of Errors across Language Conditions, and Percentage of Errors from 1080 Stimuli (%Error$_T$)*... 200

Table 46 *Distribution of Errors (Miss and False Alarm), Hits and Correct Rejections (CR) in Prediction Task, Total Number of Errors across Language Conditions, and Percentage of Errors from 720 Stimuli (%Error$_T$)*.. 201

Table 47 *Sensitivity (d') and Decision Criterion (C) across the Three Tasks (Call Sign Recognition, Error Identification, and Prediction), and Two Language Conditions, Second (English) Language (L2), and Language Switching (Mix)* 202

Table 48 *Mean Pure Response Times in Seconds (M), Standard Deviations (SD), Error Counts and Percentage of Errors within Conditions (%Error) of Call Sign Recognition (Task 1), Error Identification (Task 2), and Prediction (Task 3) Tasks with and without Background Talk*.. 214

Table 49 *Comparison of Sensitivity (D') and Decision Criterion (C) of the Three Tasks Conducted with and without Background Talk in Second (English) Language (L2), and Language Switching (Mix) Conditions*.. 217

List of Figures

Figure 1. Formulas for calculating the switch costs. 54

Figure 2. Formula for calculating the mixing costs. 55

Figure 3. Normal Q-Q plot for the outlier data point (L2, ISI 1, Similarity 2). 101

Figure 4. Normal Q-Q plot for the outlier data point (L2, ISI 3, Target). 102

Figure 5. Interaction plots for Language condition, ISI, and Similarity factors. 105

Figure 6. Profile plots for English language proficiency groups in the L2 and Mix conditions 109

Figure 7. Interaction plot for Language condition and ISI factors. 134

Figure 8. Profile plots for the English language proficiency groups in the L2 and Mix conditions. 140

Figure 9. Interaction plots for Language condition, Position, and Pattern factors. 166

Figure 10. Profile plots for the English language proficiency groups in the L2 and the Mix conditions. 168

Figure 11. Study 5 flow diagram with three consecutive tasks performed under background talk; the Call sign recognition, Error identification, and Prediction tasks. 180

Figure 16. Profile plots for Background talk factor across the three tasks (Call sign recognition, Error identification, Prediction), and two language conditions (L2, Mix). 216

List of Abbreviations

ASRS	Aviation Safety Reporting System
ATC	Air traffic control
ATCO	Air traffic controller
CAA	Civil Aviation Authority
CR	Correct rejection
CRM	Crew Resource Management
ESL	English as a second language
FA	False alarm
FAA	Federal Aviation Authority
ICAO	International Civil Aviation Organization
IELTS	International English Language Testing System
ISI	Inter-stimulus interval
L1	Native language experimental condition
L2	Second language experimental condition
LPRs	Language Proficiency Requirements
Mix	Language switching experimental condition
NES	Native English Speaking
NTSB	National Transportation Safety Board
RPDM	Recognition Primed Decision Making
RT	Response time
SA	Situation awareness
SDT	Signal Detection Theory
SNR	Speech to Noise Ratio

CHAPTER ONE

Introduction

Although there has been recent growth in the use of controller–pilot data links, which enable the exchange of text messages between air traffic controllers (ATCOs) and pilots, the primary means of communication is still air–ground voice communication over a designated radio frequency. Airspace where more than one language is used is more common than airspace where only a single language is used, and requires bilinguals to switch between the two languages. Bilinguals' language switching is the topic of this thesis, as is, to a lesser degree, the effect of bilingual environments on monolinguals who fail to understand communication in languages other than English.

1.1. Background of the Study

With the formation of the International Civil Aviation Organization (ICAO) and The Chicago Convention on Civil Aviation, in the year 1944, the English language was chosen as the official standardized language uniting all countries involved in aviation around the world. It appeared to be the logical choice of a language for international aviation, given that English had been the language predominantly used in aviation since the Second World War. Using a standardized language helps to communicate and share the same mental models of situations, and also aids the common understanding of phrases, thus helping avoid misunderstandings, one of the main causal factors involved in many air accidents and incidents (Etem & Patten, 1998).

Probably the best-known accident commonly cited in relation to misunderstanding of communication was the 1989 Avianca Flight 52 accident. As the aircraft approached its New York destination it was asked to hold over the airport due to fog. The pilot notified the ATCO that they required a 'priority landing', but because he did not use the standard terminology and declare an 'emergency', the ATCO was not aware of the urgency. The aircraft ran out of fuel and crashed, killing 73 people of the 158 aboard (National Transportation Safety Board, NTSB, 1990).

In public hearing testimony, a foreign airline captain made the point that "if a pilot, or flight crew, has a limited English language vocabulary, he has to rely heavily on the meaning of the words he does know. If those words have a vague meaning, such as the word 'priority', or if a clear set of terms and words are not used by pilots and ATCOs, confusion can occur as it did in this accident" (NTSB, 1990, p. 63). A captain from Avianca Airlines testified that the use of the [incorrect] word 'priority' rather than the [correct] word 'emergency' may have resulted from training at Boeing where 'priority' was used in procedure manuals provided by the Boeing Company to the airlines, giving the impression that the words priority and emergency conveyed the same meaning (NTSB, 1990). Yet, they do not.

Moreover, the pilot and the co-pilot of Avianca Flight 52 were speaking to each other in Spanish before the co-pilot communicated in English to the ATCO. The co-pilot switched between two languages, English and Spanish. According to Cushing (1994), the pilot told the co-pilot—in their native language—to notify the ATCO that an emergency prevailed, but the co-pilot gave the ATCO a different message in English—he told the ATCO that the aircraft was running out of fuel. Then, he told the pilot—in Spanish again—that he had told the ATCO that they were in an emergency (situation), which was not exactly what he had said in English. This is just one example that inspired ICAO to invest considerable effort in regulating the terminology by creating standardised phraseology. Its use was repeatedly emphasised.

Today, pilots and ATCOs with English as a second language (ESL) outnumber native-English-speaking (NES) aviation personnel (ICAO, 2001a); that is, communications in which at least one of the parties is an ESL speaker, dominate. Such multilingual working environments require a common language, a *lingua franca*, English. Although pilots and ATCOs are well adapted to this reality, verbal communication as the primary means for the pilot–controller interface holds many issues that remain to be explored.

Communication has been widely studied in aviation (Barshi & Farris, 2013; Cushing, 1994, 1985). As Avianca Flight 52 illustrates, there are other factors that can affect communication, such as culture, the exact translation of the meaning of one word from one language to its equivalent in another language, and the factors related to message composition. What has not been examined enough empirically, however, is the matter of alternation of two languages on the same radio frequency (or for cockpit and radio communications), and the extent to which the performance of bilingual aviation personnel is affected when they alternate between two languages.

There appears to be no empirical evidence as to whether language alternation affects the performance of bilinguals communicating in both languages used in a particular airspace, compared with communicating in one language, albeit their second language. The experience of being bilingual and operating in a mostly English-speaking environment does not appear to have been explored, nor whether there is a difference between ESL speakers living in a foreign country, and thus using English most of the time in daily interactions, and ESL speakers living in a non-English speaking country and use English infrequently, at work.

An accident near Charkhi Dadri, in India, points to the need for this type of question to be asked. In 1996, a Saudi Arabian Boeing 747 collided shortly after take-off with a Kazak Airlines aircraft on approach (Aviation Safety Network, n.d.). Aviation experts cited several factors that might have contributed to the collision, one of which was the language mix-up between the Indian ATCO and the Russian and Saudi Arabian flight crews (Orlady & Orlady, 1999; Burns, 1996). All spoke English as their second language. The Kazakhstan Airlines' pilots failed to follow the ATCO's instructions. On board their aircraft was a Kazakh radio operator, who was the *interpreter*. The radio operator did not have his own flight instrumentation but had to look over the pilots' shoulders to read the information (Rediff on the net, n.d.). The presence of an interpreter in the cockpit was a relatively frequently reported category of problems related to miscommunication in the Aviation Safety Reporting System (ASRS; Jones, 2003), as it does not allow for immediate compliance with an ATCO command. Also, when a message needs to be translated and relayed, sometimes in a hurry, the correctness of translation can be additional factor.

The general aim of this thesis was to explore whether bilinguals show performance differences between conducting tasks in monolingual settings, in their second language, and when conducting the same tasks in a bilingual, language switching environment, within an academic framework. This research aimed to investigate the cognitive processes involved in language alternation. The objective was to provide insight into potential barriers to the effective communication of bilinguals, and to determine whether the monolingual or bilingual language conditions facilitate faster response times (RTs) and fewer errors, thus providing implications for the safety and efficiency of operations.

With consideration to the general aim of this thesis, the following central research question was identified:

Are there differences in the performance of bilinguals communicating in one language compared with bilinguals switching between two languages?

1.2. Thesis Outline

Chapter 2, Study Context, begins with an overview of the thesis framework by providing the legislative background—a description of the regulations related to language in aviation and its historical roots. Chapter 3, Literature Review consists of three parts. The first part addresses the basic concepts of situation awareness (SA), which has been found to be adversely affected by the simultaneous use of two languages on the same radio frequency in several incidents and accidents. Situation awareness is approached through three underlying cognitive processes of recognition, comprehension and prediction. Then, bilingualism is explored, with a specific focus on reviewing language switching studies, and consideration of the aviation environment. The second part considers the necessary methodological assumptions that arose from the review of the literature, and the needs of the current research. A description of the Signal Detection Theory introduces the performance measures and related challenges, as well as proposed solutions. A critical review of the previous research and current knowledge in this field lead to the identification of the research problem and questions in the third part of the literature review. These suggest the need to employ an experimental approach to obtain empirical evidence of the cognitive processes underlying language alternation.

Before obtaining objective data using an experimental approach, an online survey was used to explore the current situation in aviation and verify that there were indeed difficulties with the use of two languages for radio. The current extent of the problem of language switching was questioned, with the suggestion that the situation might have changed since the implementation of the Language Proficiency Requirements. This issue was addressed first in Chapter 4, Study 1.

Chapters 5, 6, and 7 report the experimental studies, and investigate the nature of the cognitive processes of recognition, error identification and prediction. Quantitative insight into the characteristics of performance on these cognitive tasks leads to Chapter 8, Study 5, where performance on these processes while simultaneously listening to an additional source of speech, background talk, is investigated. This experiment provides a different perspective, by examining the switching between languages when attention is switched from listening to a conversation in the cockpit (or control room) back to radio communication. Chapter 9 compares these findings with those obtained without the background talk, to explore the importance of the sterile cockpit for bilinguals in a bilingual air traffic environment. The thesis is organised with a specific review of the literature for the relevant cognitive process for each study contained within that study's chapter. Chapter 10 presents a more generalized discussion of the overall research and its findings, and outlines the implications for practice and suggestions for further research, leading to the Conclusion as the final chapter of the thesis.

CHAPTER TWO

The Study Context

2.1. Bilingualism: Definition and Use

First, it is essential to define *bilingualism* and how it is used in this study. Bilingualism, as evident from the prefix (bi-), refers to an ability of an individual to use two languages (Fabbro, 1999). However, even though the communication on a radio frequency can basically be conducted in only two languages (a local language and English), aviation personnel may be multilingual—able to speak more than two languages. Colloquially, 'bilingualism' is used as a cover term to embody both bilingualism and multilingualism (Bhatia, 2017; Fabbro, 1999). This provides a certain convenience for this study to focus on bilingual language processing while acknowledging the ability of some aviation personnel to use more than two languages.

Native English Speaking (NES) aviation personnel can also be bilingual. For example, a NES pilot who also speaks Spanish would be able to understand either language spoken in airspace where both English and Spanish were used. This situation, however, will not be addressed in this thesis due to complexity of such a research design. The main reason is that, in the context of bilingual air traffic, with English as an aviation language, bilingualism usually refers to non-native English speakers with English as a Second Language (ESL). Therefore, in the context of bilingual air traffic environment and in this thesis, all aviators—whether NES or ESL—who do not understand a local language spoken in a particular part of airspace, and can use only English, will be referred to as 'monolinguals'.

Another situation that can occur but will not be specifically addressed in this thesis, is that English can be the third or even fourth language of an individual. This might be the case in multilingual countries with majority and minority, or indigenous languages (such as India, Brazil, Belgium, Luxembourg and Indonesia). For example, an ATCO can have Malayalam as their native language, Hindi as the second language, and use English as the third language. Owing to the large number of such countries, this situation is likely not rare.

However, because of the relative lack of knowledge of language alternation in aviation, it was considered both reasonable and desirable to explore the fundamental relationships before investigating the more complex issues associated with bilingualism in aviation. For readers' interest, however, there are several studies that have examined switching between native and third languages, or between two non-native languages (e.g., Costa, & Santesteban, 2004a; Gabryś-Barker, 2006; Philipp, Gade, & Koch, 2007).

To sum up, for concision, the term 'bilingualism' will be used throughout the thesis to embody bilingual ability in a native tongue, with ESL.

2.2. Bilingual Air Traffic Environment

After the establishment of English as the international language of aviation, its role was defined to be complementary to the primarily used local language of a station on the ground in documents relating to the use of language produced by regulatory authorities (ICAO, 2010; 2001). In other words, radio communications are not restricted to English only. ICAO (2001) recommends that communication shall be conducted in the language of the state and English is to be available at all control facilities serving international flights. English shall be made available when pilots are unable to use the local language of the country. Even though its role is complementary, it has a uniting function in the system. This trend persists from the early days of aviation and potentially creates a bilingual air traffic environment.

A *bilingual air traffic environment* is defined by the use of two languages, one of which is English, for conducting radiotelephony (ICAO, 2010). Specifically, it is a communication environment in which "controllers alternate between their local (usually native) language and the English language, while pilots may choose which of the available languages to use" (ICAO, 2010, 3.3.22, p. 3–7). This potentially leads to the simultaneous use of two languages in a single-frequency communication, or party-line communication.

Party line communication refers to the open radio channel, which allows pilots to hear their own clearances as well as those of the other aircraft (Hodgetts, et al., 2005). For example, in China, ATCOs will speak Chinese with Chinese-speaking airlines and English with any other international airlines. In this situation, those who can speak Chinese alternate between the two languages, while pilots who do not understand Chinese may be unable to take into account exchanges expressed in that language, which may, consequently, negatively affect their SA (ICAO, 2010; Orasanu, Fisher, & Davison, 1997). When pilots do not understand transmissions delivered in languages other than English, they may lose awareness of the position and intention of other aircraft in the airspace, because they do not receive information they might need.

Bilinguals, on the other hand, can confuse the languages of messages addressed to an English-speaking aircraft with a local language-speaking aircraft. The Avianca Flight 52 (NTSB, 1990) accident appears to show that language alternation can adversely affect not only the performance of monolinguals but also that of bilinguals, hindering their speed and accuracy of tasks performance.

2.3. Bilingualism as a Contributing Factor in Safety Occurrences

Bilingualism has been identified as a contributing factor in at least two mid-air collisions, which will be described in this chapter. When an accident happens, it is almost always the result of many contributing factors; however, because of the character and purpose of this thesis, the focus will be placed only on the linguistic factors.

Bilingualism on the same radio frequency contributed to the mid-air collision over Rio de Janeiro, in 1960, which resulted in 61 fatalities. A United States Navy aircraft was controlled in English and a Brazilian aircraft was controlled in Portuguese (Borins, 1983). The ATCO was switching between the two languages. This accident was reviewed a few years later when an initiative to convert the air traffic system from monolingual English to bilingual French and English occurred in Canada. This will be described in the next section.

Later, a similar mid-air collision occurred over Zagreb, in 1976 (e.g., Aircraft Accident Investigation Commission, AAIC, 1976; Cookson, 2009). An Inex Adria was flying from Split to Germany and a British Airways was flying from London to Istanbul. The Inex Adria aircraft was cleared to climb through flight level 330, the level at which British Airways was flying. An ATCO realised the danger of the collision and instructed the Inex Adria to stop climbing, but to do so he switched to his native Serbo-Croatian language, contrary to the regulations. This meant that even if the British Airways pilots had heard this instruction, they were unlikely to have understood it. The Inex Adria had levelled off at the flight level 330, the two aircraft collided, and 176 people died. The ATCO was found guilty; and sentenced to seven years' imprisonment. After a petition by ATCOs, the ATCO was released after serving nearly two years in prison.

The 1976 Zagreb accident was also cited by ICAO (as cited in Cookson, 2009) in relation to English language proficiency, which raises the question as to whether bilingual issues in aviation are restricted to English language proficiency per se. The Tenerife runway collision that occurred in 1977, just one year after the Zagreb accident, is known for being the deadliest accident in the history of aviation. It can illustrate a deeply rooted problem beyond a discussion of English proficiency in aviation. Besides many other contributing factors, the linguistic problem was in the utterance "We are now at take-off", made by the KLM captain speaking to a Spanish ATCO (Tajima, 2004). This is a non-standard phrase that could be interpreted as either "We are now at the take-off position" or "We are now taking off." Indeed, the ATCO interpreted the phrase to mean that the pilots were waiting for a take-off clearance because he instructed them to "stand by for take-off"—to wait for a further take-off clearance. His transmission was then interrupted by someone else's transmission. In the meantime, the KLM pilots had already started their take-off run, colliding with a Pan Am aircraft just a moment later.

Subsequent analyses explained that in Dutch, the present progressive tense of a verb is expressed by the equivalent of the preposition "at" in English plus the infinitive of the verb (Tajima, 2004). This might be what the KLM captain did. He unintentionally used the present progressive form of his native Dutch language when he spoke English, indicating that what he really meant was that he was taking off (Tajima, 2004). The captain of the KLM aircraft was the chief examiner and an experienced pilot, both in flying and in using English. The Netherlands is one of the European countries that had permitted only English for its air traffic communications (Tajima, 2004). This suggests that language-related errors are not simply the result of an insufficient command of English. This can bring some doubt into whether the effects of bilingualism on the performance of bilinguals have been sufficiently explored and understood.

2.4. Bilingual Air Traffic Conflict

While some countries have considered converting to monolingual English air traffic systems (e.g., China), Canada converted to a bilingual French and English system in 1979 after a conflict of more than ten years. It started as an initiative of francophone pilots in 1962. At this time, a flying school opened in a remote part of Canada, which dominated by French-speaking pilots. The pilots were trained for their private pilot licenses and flew only within the region's visual flight rules. Therefore, it was not considered necessary to require them to use English for radio communications. Subsequently, permission to conduct radio communication in French was requested (Borins, 1983).

A memo was issued by the Director of Civil Aviation in the same year allowing ATCOs to use the French language in conditions of emergency or stress and requiring ATCOs to translate the message into English as well, so that every pilot in the vicinity would comply. The memo was confidential for several reasons. In particular, if the information had spread among the pilots, the use of the French language on a radio frequency would become more and more frequent, and this was not desired. The aim was only to permit its use under extreme circumstances, but not to encourage its use (Borins, 1983).

Initially, ATCO positions were dominated by English-speaking personnel. This led to an initiative and changes in recruitment policies for French-speaking candidates. For example, if a French-speaking candidate passed a test of three-dimensional abilities, but failed in other tests, they were admitted to the programme provisionally, with a requirement for English language training and an examination that they had to pass before starting practical training (Borins, 1983). As the number of French-speaking ATCOs grew, a debate over whether the language of aviation should be English or French began. A factor that favoured bilingual air traffic was the acceptance of the Official Language Act by the Canadian government in 1969, a law that made both French and English official languages of the country (The Canadian Encyclopedia, n.d.).

The controversy of opinions prompted the pilots' and ATCOs' strike against bilingualism in 1976 (Borins, 1983). Because the elementary facts in the dispute were unavailable, appointed representatives conducted visitor trips to airports in European countries where bilingual air traffic control (ATC) was used, such as Charles de Gaulle Airport, France.

More importantly, there was an effort to find empirical evidence that would shed light onto the debated arguments. Three empirical studies were conducted to compare the performance of bilinguals between monolingual and bilingual air traffic conditions, which also influenced the development of this thesis, given that no one had ever before carefully studied whether one system is safer than another (Borins, 1983). No further studies have been completed. These studies will be described in detail in section 3.4.

Many other factors, including politics and media reporting, played a role in the resolution of the conflict, which, at its beginning in 1962, looked to favour a monolingual English operation, yet resulted in a bilingual system in 1979. In fact, it was just three years after the Zagreb accident that a bilingual ATC system was acknowledged as safe by the Canadian government, resulting in conversion from a monolingual ATC system with the use of English only, to a bilingual, French and English ATC system. The official accident report on the Zagreb accident was issued by the AAIC in 1976, and could have potentially influenced this decision, but seems to have escape the attention of the Commission, which drew no firm conclusion about the accident. The Commission considered the key benefit of bilingual ATC to be "the increased comfort and safety for francophone pilots, who would be able to use their mother tongue" (Borins, 1983, p. 187).

The case of Canada is probably the only example of conversion from a monolingual to a bilingual air traffic system (Borins, 1983). A misperception might occur as to the extent to which ICAO was influenced by the bilingual air traffic conflict in Quebec when considering the possibility of using two languages for radio broadcast, given that its headquarters are in the heart of Quebec. However, ICAO's provision about languages became effective in 1950 (ICAO, 2001a), well before Canadian civil aviation converted to a bilingual air traffic system.

Communication was one of the first types of aviation procedure to be standardized as a solid foundation of post-war civil aviation (Borins, 1983). The Chicago Convention (The Convention on International Civil Aviation) was signed by 52 states in 1944 (ICAO, n.d.a). ICAO was established the following year, first as the Provisional International Civil Aviation Organization (PICAO), and then, in 1947, as it is known today (ICAO, n.d.a). Right from the beginning, ICAO recognized that many countries would like to use their own language in addition to English (Borins, 1983). In response to this, a broader context of the language-related aviation legislation is explained in the next section.

2.5. English Language Proficiency Regulation in Aviation

ICAO recognised that the establishment of a single language in radiotelephony would face several challenges, which are described in Document 9835 (ICAO, 2010). Among the concerns was that implementation of a single language policy would exclude many currently active pilots and ATCOs due to their limited English language proficiency.

Probably one of the most striking accidents to illustrate the need for improvement of English language proficiency among aviation personnel was the fatal air crash of a China Northern Airlines aircraft at Urumqi, China, in 1993. The aircraft descended too steeply, but the crew did not notice it because of fog. When the ground-proximity warning system gave its alarm, one crew member asked the other, in Chinese, what the words “pull up, pull up” meant. The aircraft crashed shortly afterward while they were discussing the meaning of this phrase in their native language (Orlady & Orlady, 1999). In 1995, insufficient English language ability led to ambiguities in communications between a Spanish-speaking ATCO and the English-speaking crew of an American Airlines flight, which contributed to a crash into a Colombian mountain. The ATCO, however, did not communicate that some of the crew’s reports and requests were not understood (Simmon, 1998).

In response to the need to address the issues related to language, ICAO started an initiative for English language proficiency improvement to enhance aviation safety (ICAO, 2010). In 1998, ICAO issued a policy of English Language Proficiency Requirements (LPRs), which addresses personnel licensing related to English proficiency. All aviation personnel shall obtain at least the minimum level specified, Operational Level 4—considered the safety threshold—of the six-point rating scale. The requirements were implemented in 2011, three years after its application date of 5 March, 2008 (ICAO, 2010).

English language ability is tested according to six criteria: pronunciation (which requires the person’s accent to be intelligible to the aeronautical community), structure (relevant grammatical structures and sentence patterns), vocabulary range, fluency of speech, comprehension, and interaction (ICAO, 2010). To achieve test quality, ICAO recognized the need for the assessment of the English tests to conform with the LPRs, meaning a test must meet established criteria and follow good testing practices (ICAO, n.d.b). The good testing practices, however, were not further defined in this particular source of literature (ICAO, n.d.b). In 2011, ICAO launched the Aviation English Language Test Service website to support the entire process, with qualified experts providing detailed guidance on the test assessment service (ICAO, 2013).

Seven years have passed since the implementation of LPRs, so it is timely to consider what has changed since then and whether the challenges are still the same. The following section addresses the situation after the implementation of ICAO's LPRs.

2.6. Language Issues in Aviation after ICAO LPR Implementation

Problems related to bilingual air traffic system are not just historical issues. The use of two languages for radio communications was identified as one of the contributory factors in a fatal runway incursion at Paris Charles de Gaulle Airport on May 25, 2000 (Bureau d'Enquêtes et d'Analyses, BEA, 2000). The crew of the UK carrier Streamline Aviation were probably not aware that the Air Liberté was going to take off as its clearance was issued in French (BEA, 2000). The two aircraft collided and the co-pilot of the Streamline was killed when the left wing of the Air Liberté collided with the right propeller of the Streamline and cut through the cockpit (BEA, 2000). Consequently, French air accident investigators recommended that English be used for all ATC communications at major airports in France (BEA, 2000). However, Alcock (2007, para. 3) indicated on the Aviation International News website that "French pilots have resisted pressure for it to be made mandatory in France." Recordings of air traffic communications in Paris demonstrate that ATC communications there are still bilingual (Gaëtan, 2016).

In 2001, Milano Linate airport, Italy, experienced an accident between a departing Boeing MD87, operated by Scandinavian SAS, and a German-operated Cessna Citation C525 taxiing for departure, which captured the headlines of the New York Times and the Guardian (Henneberger, 2001; Willan, 2001). All 114 people on board the two aircraft, and four ground personnel were killed. The visibility was low and traffic levels were high.

ATCOs on duty were assisting 24 aircraft, using both English and Italian on the radio frequency (Agenzia Nazionale per la Sicurezza del Volo, 2004). Importantly, bilingualism was not the primary cause of the accident—the pilots of both the aircraft were speaking English with the ATCOs. This can serve to stress that when an accident or incident occurs in a bilingual air traffic environment, it does not necessarily mean that this particular factor contributed to its occurrence. However, the fact that bilingualism on a radio frequency is seldom one of the main causes in the error chain of events leading to an accident, can lead to bilingualism escaping the attention of thorough research analyses. This indicates the need to empirically clarify the actual effect of bilingualism in radio communications on the performance of bilinguals.

Some initiatives have been taken to address the phenomenon of bilingualism. In 2006, the Air–Ground Communication Safety improvement initiative was launched by EUROCONTROL. The report had a wide scope with language being a small part of it. Two years later, Prinzo and Campbell (2008) started a series of interviews with US pilots regarding their difficulties in international operations, in order to study the problems related to bilingual air traffic. Prinzo et al. (2010a, 2010b) found that pilots perceived increased workload in bilingual situations, which was attributed to their attention being focused on non-English transmissions in party line communications. The studies mentioned here summarised the recommendations made by pilots and ATCOs, suggesting the use of only English for international general aviation operations.

Presumably, English language proficiency improved after the implementation of LPRs in 2011. For example, in 2010, at St. Petersburg, a Russian ATCO did not understand messages from a Swiss Airbus, which needed to return to the airport after a bird strike. A pilot from another aircraft translated the messages into the ATCO’s native language (Mori, 2010). A year after the LPR’s were implemented, in 2012, bilingualism contributed to a loss of SA (two aircraft were not aware of each other) in Spain, when an ATCO spoke English to Brussels Airlines A319 and Spanish to Iberia A330-200. The incident resulted in a loss of separation on final approach to Barcelona (Hradecky, 2013). Similar to the recommendation made in France after the fatal runway incursion (BEA, 2000), the Spanish Airports and Air Navigation provider stressed “the importance of using the English language when it is spoken by any aircraft crew” in the final report released by Spain’s Civil Aviation Accident and Incident Investigation Commission (2013, p. 218).

According to Yun (as cited in Dennis, 2015b), an official at the Civil Aviation Authority of China (CAAC), China has considered the use of a monolingual English air traffic system. At present, ATC in China uses English when communicating with foreign airline pilots but Chinese (Mandarin dialect) with pilots of Chinese airlines (Dennis, 2015b). Foreign airlines flying into China want ATCOs to communicate only in English so that all pilots share common SA of the airport environment (Dennis, 2015a). Yun responded that the CAAC "took the feedback from the pilots seriously to make the decision for a common language to be used" (as cited in Dennis, 2015b, para. 6) and stated that "China will mandate that its air traffic controllers use only English… starting in 2017" (as cited in Dennis, 2015b, para. 1). This information could not be clarified with the CAAC, and thus it remains unknown as to whether this was successfully implemented. Nonetheless, this "step forward for China's aviation industry" (as cited in Dennis, 2015b, para. 5), as described by Yun, should "improve situation awareness for foreign pilots" (as cited in Dennis, 2015b, para. 2).

An example of a creative approach to the improvement in English language proficiency lies in a strategy of training non-native-English speaking pilots in English speaking countries. For example, Chinese commercial pilots train in the US (Hall, 2017; Xinhua News Agency, 2007), but first have to meet a requirement to be proficient in English (Hall, 2017). For example, in 2009, some safety issues occurred in Australia due to the lack of English language proficiency of foreign student pilots (Estival & Molesworth, 2012).

Despite efforts to improve English language proficiency, according to Mathews (as cited in Grady, 2017, para. 1), "language issues in aviation are not investigated as thoroughly as other factors", and the role of language as a contributing factor in accidents is underestimated. Indeed, there seems to be no empirical evidence as to whether bilingualism is more pragmatic and preferable for bilinguals who can communicate in both the languages used in a particular piece of airspace than speaking in only one language. Although the ability to use the native language might seem to be an advantage (Borins, 1983), there is a gap in the existing knowledge regarding the effects of language alternation on the SA of bilinguals.

To sum up, two terms used in this thesis were defined: 'bilinguals' refers to those aviators who can use both English and a language of a non-English speaking country, and 'monolinguals' refers to aviators who can use only English for air traffic communications.

Since the implementation of English as an aviation language and the acceptance of the use of another language for radio communications, insufficient English language proficiency of bilinguals has been repeatedly identified as a factor in a number of aviation accidents or incidents. Research initiatives have, however, focused mostly on the perspective and experiences of native-English speaking pilots operating in bilingual air traffic environments. Typically, an adverse effect of bilingualism on their SA was identified. There was some indication that the SA of bilinguals was also adversely affected. Effort has been put into the improvement of the issues related to bilingualism, yet the effect and efficiency of bilingual operation on performance of bilinguals does not appear to have been analysed.

CHAPTER THREE

Literature Review

3.1. Overview

The following three sections of the literature review consider the current knowledge on bilingualism. The first part defines the three key concepts identified in previous chapters, namely, situation awareness (SA), bilingualism (as cognition), and language alternation as a specific aspect of bilingualism. These give rise to the need for critical methodological considerations, which are reviewed in the second part. The third part summarises aspects of the literature review, proposes the research problem, and presents the research questions.

Part I: Defining the Concepts

3.2. The Importance of SA in Aviation

The previous chapters have identified bilingual air traffic communication as one of the contributory factors to aviation incidents and accidents, by impairing the SA of a crew (e.g., Prinzo et al., 2010a, 2010b; EUROCONTROL, 2006). According to the Civil Aviation Authority (CAA) in the United Kingdom (CAA, 2014), the loss of SA can be easy to identify in hindsight, as a conclusion of an accident investigation; however, there is a lack of analysis of causes that contributed to its loss. To investigate the causal effects, an empirical analysis must be conducted. To study SA empirically, the concept has to be precisely defined. However, “a commonly accepted definition is still missing” (Sarter & Woods, 1991, p. 45; see also Orlady & Orlady. 1999, p. 256). Additionally, it is generally accepted that the objective measurement and analysis of SA is difficult (CAA, 2014). Consideration of approaches to the definition of SA in aviation and the issues related to its measurement—such as adequately controlling for variables so that the findings can clearly be attributed to the language switching factor—has led the focus of the research to a cognitive perspective.

According to the CAA (2014), SA has been studied in aviation since the 1990s. It was either understood as a cognitive process, or a variety of cognitive processing activities for building and maintaining awareness of a situation or event (e.g., Bryant, et al. 2004; Flin, O'Connor, & Crichton, 2008; Sarter & Woods, 1995), or was not even considered as a psychological construct (Patrick & James, 2004). SA is commonly understood as "knowing what is going on around" (CAA, 2014, p. 73). Endsley (1995) understood SA as a state of knowledge. In the most commonly cited definition proposed by Endsley (1995), SA entails three levels: perception, comprehension and recognition.

SA and information processing can be used either interchangeably (CAA, 2014), or as two separate concepts (Rousseau, Tremblay, & Breton, 2004). For example, Endsley (1995) claimed that SA is *a state* of knowledge that, in fact, needs to be distinguished from the *processes* used to achieve that state. Specifically, situation assessment refers to a process of obtaining SA, and SA refers to a product of the situation assessment; that is, a state of knowledge (e.g., Billings, 1995; Sarter & Woods, 1991, 1995). Rousseau, Tremblay and Breton (2004) added that SA cannot simply be equated to any verbal report of the status of consciousness about a situation, which makes the measurement of SA in the context of language perception—the main concern of this study—a big challenge. It was also argued that the phenomenon of SA should be explained by established psychological constructs that have an identity separate from the phenomenon itself (Shebilske, Goettl, & Garland, 2000).

Durso and Gronlund's (1999) distinction between two approaches to SA may help to reconcile the different definitions described above. They distinguished between the *operator-focused* approach and the *situation-focused* approach. The former entails the mechanisms on the side of an operator that determine SA, and the latter addresses the determinants of the environment and situation in which an operator works. Indeed, the term SA does not only consist of awareness as a cognitive process or concept, but it is also situation dependent. A state-oriented definition could, therefore, be associated with a situation-focused approach (Rousseau, Tremblay, & Breton, 2004), meaning that SA modelling and analysis is driven by the elements of a situation, and would therefore vary from situation to situation. Orlady and Orlady (1999) distinguished five types of situations in which SA is involved:

1. Aircraft mode and status SA refers to attention to cockpit dials, gauges or instruments, indicating, for example, flap position, power used or flight profile, and mode of the automatic flight control system.
2. Awareness of aircraft position in respect to the flight track, obstructions and other aircraft relates to navigation tasks. Place information can be verified by visual contact, by attention to maps, charts, the flight plan, radio and the cockpit instruments. This is especially important during IFR[1] flights.
3. Awareness of the external operating environment relates to external conditions, such as weather, airport infrastructure, ATC and radio communication. Meteorological phenomena such as wind shear, ice, snow, and thunderstorms, and existing and forecast weather are important considerations for operation.
4. Awareness of the mental state of other team members (the cockpit and cabin crew) and passengers. The state of mind, behaviour or fatigue of others can contribute to safety. Probably the most alarming example of this category, which is seldom addressed in research, is Germanwings Flight 9525 in 2015, where the first officer deliberately flew into the French Alps (BEA, 2016).
5. Element of time. The concept of SA has a temporal dimension, as it requires pilots to think ahead of the aircraft. For example, it is important to be aware of the fuel status so that the aircraft does not run out of fuel before reaching the destination airport. The temporal factor is also importantly associated with weather forecasts, and with flight planning (time over planned navigational points).

The situation-focused approach limits the description of cognitive processes involved in SA (Rousseau, Tremblay, & Breton, 2004). Processing of language, however, is a cognitive process not specifically dependent upon a situation. When considering SA in relation to language alternation, it is not the particular situation that is of relevance, but rather the operator’s perception and processing of verbal information from which SA can be constructed. According to Rousseau, Tremblay, and Breton (2004, p. 6), “if one is to improve SA, the elements of the situation critical for SA should be specified, and the SA content definition should follow from these elements. On the other hand, if SA depends on

[1] Instrument Flight Rules – “Rules and regulations… to govern flight under conditions in which flight by outside visual reference is not safe. IFR flight depends upon flying by reference to instruments in the flight deck, and navigation is accomplished by reference to electronic signals.” (Federal Aviation Administration, FAA, 2012, p. G-9)

a set of processes that are not an intrinsic part of SA as a state but on which SA depends, it becomes important to specify which processes are essential to SA. SA improvement, for instance, will depend upon changes in the operation of these processes."

A process-oriented definition can be associated with an operator-focused approach, and thus would examine characteristics of an operator, especially the set of cognitive processes supporting the production of the mental representation corresponding to the SA state (Rousseau, Tremblay, & Breton, 2004), such as attention and memory, with which SA is commonly associated with (Endsley, 2000). According to Endsley's model of SA in dynamic decision making (1995), memory (and other individual factors) affects SA, and SA affects decision making. In other words, decisions are made on the basis of a person's SA, and memory influences the entire process (Endsley, 1995). For example, when taking actions in a dynamic system, such as aviation, an operator perceives and processes the elements in the environment, but this might be adversely affected by the limits of the short-term memory. Similarly, it seems reasonable to assume that information and experience stored in long-term memory might further affect the decisions about the situationally appropriate action. Even Endsley's perception, comprehension and projection corresponding to the SA state definition are basically cognitive processes (Rousseau, Tremblay, & Breton, 2004). That is, they highlight that an explicit description of the processes involved in providing operators with cognition is required. However, Bryant, et al. (2004, p. 107) argues that Endsley's model does not allow "detailed specification of the cognitive operations underlying SA", but "it has delineated classes of cognitive processes linked to errors of performance." Bryant, et al. (2004, p. 105) postulated that although Endsley's model is descriptive, attempting to provide "a broad overview of the different stages that information passes through before decision is made", it is not specific enough to provide a prescriptive model that can "identify the rules and heuristics that guide the operations within and between each stage of processing". Nevertheless, "knowledge of the processing architecture is essential to understanding how different kinds of information are treated" (Bryant, et al., 2004, p. 105).

Endsley's first level of SA, perception of the elements of a situation, was considered equivalent to recognition of what is happening at the time (CAA, 2014). Errors, then, represent failures to perceive information, probably caused by a lack of detectability of sensory data (Bryant, et al., 2004). This directs attention to the Signal Detection Theory

(SDT; described in section 3.10) and advocates its use in the research methodology. Endsley's second level of SA addresses the errors related to understanding the meaning and significance of what was perceived. In other words, not only what it currently means, but what implications it has for the subsequent development of the situation (CAA, 2014). Processing the dynamics of a situation leads to Endsley's third level of SA, the projection of future status, based on an assumption of what will happen next. Subsequent decision-making for what needs to be done or is appropriate to do (CAA, 2014) can be considered the result of the process. This would suggest that experience and subsequent automaticity are important factors, both of which might significantly benefit SA (Endsley, 1995). However, Endsley (1995, p. 45) argues that it is questionable "to what degree do people who are functioning automatically have SA", provided that awareness implies consciousness of the information. Furthermore, an operator can be situationally aware even in a novel situation with which the operator has no prior experience. Thus, it can be assumed that while SA is not necessarily dependent on experience, it can potentially be nurtured by practice.

The largest number of errors identified from accident and incident analysis appear to have occurred at the first level of recognition (Endsley, 2000). Jones and Endsley (1996) analysed 143 SA-related incidents and classified 76.3% of errors as level 1, 20.3% as level 2, and only 3.4% as level 3. Examples of how the three levels of SA can be associated with language include:

1. The perception of important elements (e.g., call sign recognition in bilingual air traffic environments)
2. The comprehension of the meaning (e.g., what does the ATCO mean by that? Was there a mistake in the provided information?)
3. The projection of the future state (e.g., if the situation follows this pattern for a certain time, will any potential hazard appear?)

SA is a cognitively complex subject (Endsley, 2000) and can be involved in many different situations beyond the bilingual air traffic environment. However, the focus of this research is on the bilingual air traffic environment and its potential effects on SA acquisition of bilinguals. Even if radio communications were conducted only in English, some kind of issues related with dual language processing will likely still persist, given that the majority of aviation personnel are non-native English speakers (ICAO, 2001a). For example, a crew can communicate in their native language when they are not communicating by radio (e.g., “How are the children”, in Chinese), so when they need to resume speaking and listening in English when continuing radio transmissions, they must switch between their native language and English. It is therefore assumed that, from a cognitive perspective, exploring SA in the context of language alternation might bring more beneficial overview of aviators’ cognitions from operating in a bilingual environment than from the simulation of specific situations. If SA is situation dependent, the cognitive approach is critical, because it allows the generalization of findings beyond the target scenario, which situation-based approaches may not.

To sum up, focusing on selected cognitive correlates or aspects of SA allows for the choice and use of direct performance measures (i.e., performance speed and accuracy) despite its cognitive complexity. This could be a potentially fruitful approach to the measurement of the effects of language alternation on performance, particularly in determining whether there is a relationship between language alternation and SA. A cognitive, operator-focused research approach can provide helpful empirical evidence of correlations between the bilinguals’ language alternation and their SA.

Based on the discussion above, the following *requirements* for the development of the methodology were formulated to allow systematic exploration of the topic:

1. The operator-focused approach to SA should be employed to measure the three levels of SA through the underlying cognitive processes of recognition, comprehension, and prediction.
2. The three cognitive processes should be investigated individually by developing three experimental tasks.
3. A task designed to investigate a higher level of SA should encompass the level (or levels) below it. In other words, a task examining comprehension should continue using the same type of acoustic stimuli used for the analysis of recognition, because it is reliant on it (CAA, 2014). Yet, the task will be designed to test comprehension instead of simple recognition, and as such, would be different. Similarly, the third task, testing prediction, should continue using the same type of acoustic stimuli as the previous two experimental tasks.
4. Consequently, a measurement of RT must be considered to provide objective information reflecting a particular level of SA.

5. Given that SA may be affected by distraction (CAA, 2014), all three tasks shall also be examined under the condition of distraction.

3.3. Bilingualism, SA and Cognition

Both SA and bilingualism have long been of interest (e.g., Hoffman, 2015; Keatley, 1992; Pavlenko, 2014). However, there seems to be a lack of research on the relationship between SA and bilingual communication. Therefore, the nature of bilingual processing in situations requiring awareness remain unclear. In contrast, the nature of language alternation and cognition has long been studied in psycholinguistic, and neural studies (e.g., Costa, Miozzo, & Caramazza, 1999; Hernandez, et al., 2010; Macizo, Bajo, & Paolieri, 2012; Martin, Macizo, & Bajo, 2010). However, the psycholinguistic studies were not conducted with the aim of addressing any aviation issues and therefore are not directly relevant to aviation. This section considers the overlap of aviation SA and bilingualism.

Bridging the two constructs might be facilitated by a cognitive approach. Approaching SA through its underlying cognitive processes, as was suggested in the previous section, would allow exploration of the relationship between its three cognitive processes and bilingualism. Specifically, bilingualism turns attention towards language processing; and language belongs to one of the critical areas of research in cognitive psychology (Kellogg, 2016). Therefore, cognition might help explain the overlap of SA and language.

Language can be defined as a system of symbols used to communicate ideas among two or more individuals, implying that, in conversation, the speaker and listener exchange mental representations using verbal symbols (Kellogg, 2016). Additionally, the simplest notion of SA is when perception matches reality (Air Force Flight Standards Agency, AFFSA, 1998). When SA involves integration of information received through the sensory channels over a period of time (AFFSA, 1998), it can be said that using correct spoken symbols to communicate ideas between pilot and ATCO further determines their SA. In another words, adequate understanding of communicated mental representations of an ATCO, who uses spoken symbols (language) affects the pilot's SA, and vice versa. Ideally, an individual shall first correctly encode a mental representation into words and use proper language. Therefore, it can be said that the encoding process can determine the effectiveness of communication. Unfortunately, encoding can be prone to mistakes as what is meant is not always precisely expressed with words ("I thought I said…/I meant…"). Besides the accuracy of the operator's encoding of a message, there is another factor, which relates to language itself, that can determine SA.

Ambiguity of utterances is a factor which significantly decreases SA. It can be caused by the use of homophones (e.g., brake/break, two/to, missed/mist, hear/here), homographs (e.g., the word 'bear' can mean an animal or to incline in an indicated direction; 'wind' can mean to turn or can refer to moving air; 'close' can mean nearby or the opposite of open), and homonyms (e.g., 'left' can be a direction or can mean departed), which are abundant in the English language and can potentially cause confusion, especially for non-native English speakers (Jones, 2003). For this reason, a standardized form of aviation language was developed, containing simple, unambiguous words with well understood meanings. This is known as standard phraseology (ICAO, 2010), and allowed everyone, especially those with a more limited command of English, to be situationally aware.

The standardized form of communication is suitable for routine situations, which allow using the reduced vocabulary of around 400 words with a precise meaning (e.g., "Roger" means "We have received all of your last transmission, understood it and we shall proceed as expected"), and short sentences (e.g., "Say again") (ICAO, 2010). However, in situations of distress and in non-routine situations, communication requires the use of plain language, "the spontaneous, creative and non-coded use of a given natural language" (ICAO, 2010, p. 3-5). The plain language register puts larger demands on the English language proficiency of non-native English speakers, but also on adjusting the speech of native-English Speaking (NES) aviators (Jones, 2003), as it includes the use of a wider vocabulary (often with less precision and agreement about the meaning of the terms), topics, and, longer and less organized sentences (ICAO, 2010).

In addition, according to Jones (2003), approximately 49 instances of different terminology are used in the United States and elsewhere, given by the discrepancies in the documents produced by regulatory authorities, the American Federal Aviation Authority (FAA) and ICAO. For example, ICAO uses 'stop' and the FAA uses 'hold'; for moving away from something, ICAO uses 'vacate' and the FAA uses 'exit' (Jones, 2003). These can represent additional opportunities for confusion along with the 44 different definitions for aviation words employed by the FAA, despite the inter-governmental committees whose mission is to unify the expressions used in aviation communication (Jones, 2003). Being situationally aware, as suggested, not only means knowing a language but also being aware of these differences in standard phraseologies to be able to communicate and act accordingly.

An additional factor can be recognized when operators perceive speech, which can be crucial in the process of building and maintaining SA. Campbell and Bagshaw (2002, p. 143) noted that "an individual only actually listens to about a third of what is heard." This was attributed to the *attention loop* (Campbell & Bagshaw, 2002), meaning that the first part of the communication is listened to, the information is evaluated, and a plan of action is formulated. This suggests that elements of the message can be missed because of the limitations of human perception, regardless of the intentions of the operator. If this phenomenon is real and appears regardless of whether the language used is native or second, it raises the question of whether there is a difference in language perception when listening to two different languages that alternate. Language control is necessary to ensure that bilinguals stay within the target language and select the correct mental lexicon—where knowledge about the meaning, pronunciation, and syntactic characteristics is stored (Aitchison, 1994)—in order to attribute the correct meaning to a perceived word or sentence (Declerck, Koch, & Philipp, 2015).

Three basic functions are involved in *spoken word recognition*: lexical access, lexical selection, and lexical integration (Zwitserlood, 1998). *Lexical access* concerns the relationship between the sensory input and the mental lexicon. Perceptual parsing processes extract information relevant for the lexical system based on the speech input. *Lexical selection* involves recognition. From the subset of activated elements and possible words, the one that best matches the speech input is selected. *Lexical integration* concerns the binding of syntactic and semantic information associated with words into a semantic and syntactic representation of the whole sentence. Because of the cognitive complexity of SA within bilingualism, this type of microscopic model and the definitions of terms might add limited explanatory value. This raises the need to move away from testing individual language control models to testing language alternation within a SA related task.

Grammar has also been recognized as a factor in individuals coming to somewhat different views of the observed world. Whorf (1940, cited in Pavlenko, 2014, p. 10) indicates that “users of markedly different grammars are pointed by their grammars towards different types of observations and different evaluations of extremely similar acts of observation.” This suggests that it is through language that a provisional analysis of reality is made. The Tenerife accident gives an example of a fatal misunderstanding due to the use of the two different sets of grammar rules (Tajima, 2004). It may therefore be interesting to explore whether being bilingual can also facilitate the process of building SA, given that the use of more than one lexicon allows for corrections to the provisional analyses of reality, through which SA is constructed. The different grammar rules of Chinese and English language will be addressed in more detail in an upcoming section, to provide the necessary understanding for subsequent stimuli development.

That bilinguals can choose between two languages led researchers to investigate whether using more languages exacts a cognitive ‘price’. However, many studies (e.g., Bialystok, 2010, 2009; Christoffels, Kroll, & Bajo, 2013; Moreno et al., 2010) demonstrate that this is not the case. Bilingualism seems to give some *cognitive advantages*, such as better selective attention (e.g., Chung-Fat-Yim, et al., 2017; Costa, et al., 2009; Costa, Santesteban, & Ivanova, 2006; Friesen, et al., 2015; Tao, et al., 2011), better conflict monitoring (Costa et al., 2009; Costa, Hernández, & Sebastián-Gallés, 2006) or greater thinking flexibility (e.g., Adi-Japha, Berberich-Artzi, & Libnawi, 2010; Bialystok, Craik, & Luk, 2012; Ibrahim, Shoshani, Prior, & Share, 2013). All three of these advantages are potentially very important and relevant to SA and operating in the aviation working environment.

The cognitive advantages of bilingualism might suggest that a bilingual air traffic environment could even provide bilinguals with some benefits in terms of task performance. However, Bhatia and Ritchie (2013) emphasized that strong conclusions about bilingualism and cognition are not warranted and the only potential benefit of bilingualism is language choice. Bilingual air traffic simulation studies, discussed in the next section, attempt to provide some empirical insight into this issue.

3.4. Bilingual IFR Communications Simulation Studies

Bilingual IFR communications simulations studies were conducted in Canada in an attempt to resolve the bilingual air traffic conflict (Borins, 1983). The simulation exercises sought to compare performance between bilingual and monolingual IFR ATC under the most realistic representation of the real world possible (Borins, 1983).

The simulations were run in three phases; an *en* route ATC, terminal ATC (studying listening watch), and terminal ATC in exceptional situations, such as bad weather conditions (Borins, 1983). The procedure of the Bilingual IFR Communications Simulations Studies were as follows. Four ATCOs sat in one room and in another room were pilots or clerks trained to copy the behaviour of actual pilots. A pilot group conversed with ATCOs and entered information into a computer, which simulated radar images of the aircraft they were flying (Borins, 1983, p.192). Almost 60 aircraft could be shown on the system at a given time, depending on the task, which is well above average air traffic levels. Testing was run in three-day periods including an orientation day. On the second day, the exercises were conducted under monolingual traffic conditions, and on the last day, similar exercises were performed under bilingual conditions; that is, with 25–35% of the pilots speaking French. Due to technical shortcomings of the experiments, the ATCOs kept track of the French language aircraft by designating their flight progress strips with yellow marking pens. Each day ended after a debriefing session during which a conflict of opinions often occurred, because some participants supported bilingual air traffic and some were its opponents. However, the debriefing provided opportunities for the participants to modify their attitudes (Borins, 1983).

There were three dependent variables in the study: the performance speed (RT), accuracy (number of errors) and losses of separation. The duration of the ATCOs' messages, on which a RT was measured, was recorded as well. The results of the simulation exercises suggested that average ATCO transmissions were slightly longer in French than in English in both the *en* route and terminal exercises (Borins, 1983). More importantly, the number of errors on the bilingual days was 8% higher than on monolingual days (Borins, 1983). The errors were differentiated on false starts, in which a transmission started in the wrong language but was corrected before it was finished, and language changes, in which the entire message was in the wrong language for that particular aircraft. The latter type of error occurred less frequently (Borins, 1983). Despite 8% more errors in the bilingual condition, it was argued that “more experience and better procedures would be able to reduce language errors” (Borins, 1983, p.186). The participants, however, were concerned and considered the number of errors to be high.

There was no statistically significant difference in the number of aircraft separation losses between monolingual and bilingual days. Indeed, the number of losses of separations was almost equally divided between the two days, monolingual and bilingual, in all three phases of the simulation exercises. The study concluded that “language was not found to be the cause of any separation loss” (Borins, 1983, p. 203). This seems to be in contrast to the accidents that were discussed in sections 2.3 and 2.6. Even though the use of language other than English was not the primary cause of the described accidents, the accidents occurred in bilingual environments. Yet, based on the results from Borins’ simulator studies, it is difficult to judge the extent to which bilingualism contributed to their occurrence. The conclusion of the study was that no particular differences between monolingual and bilingual days were found and, therefore, bilingual operations would cause no loss in system efficiency. However, it remains unclear whether bilingualism and impaired performance, such as losses of separation, are truly unrelated.

The preceding discussion regarding the advantages of bilingualism and findings that performance on bilingual days elicited more errors opens up another question. Is there a difference between cognitive gains and cognitive demands associated with bilingualism on a radio frequency? The performance difference might have been due to the language alternation putting larger demands on cognitive processing, or it might, as the author suggested, have been the lack of practice that affected performance. Like any other cognitive process, the process of language alternation is not directly observable (Zwitserlood, 1998), and therefore focus on cognitive control during dual language processing is required to complement the complex real-life simulation observations, to answer this question.

3.5. Language Switching

It has been suggested (e.g., Hanulova, Davidson, & Indefrey, 2011; Hermans, et al., 2011; Macizo, Bajo, & Paolieri, 2012; Rodriguez-Fornells, et al., 2005) that both native and second languages are still active in the brain and, therefore compete for selection every time bilinguals are about to speak in either of the languages. Both listening and speaking require cognitive control, called language control (Declerck, Koch, & Philipp, 2015; Declerck & Philipp, 2015a; Green, 1998), which is defined as a process that ensures that bilinguals communicate in a target language (Declerck, Koch, & Philipp, 2015) while a non-shared language is inhibited (Rayner & Ellis, 2007). Therefore, language control guides lexical selection (see section 3.3); that is, the process of spoken word recognition (Declerck & Philipp, 2015a).

Typically, to investigate the underlying mechanism of language control, an experimental approach called *language switching* is employed (Declerck, Koch, & Philipp, 2015; Declerck & Philipp, 2015a; Green, 1998). Here, two terms must be distinguished—— language switching and code switching. In the ICAO document 9835, the term code switching is used to refer to "the alternation between two or more languages, dialects or registers in a single conversation or a single sentence within a conversation" (ICAO, 2010, p. 3-6). In the psycholinguistic literature, language and code switching are differentiated (Declerck, & Philipp, 2015b; Lei, Akama, & Murphy, 2014). Code switching refers to the synchronous concomitant use of two languages as targets of simultaneous translation, and language switching requires diachronically parallel use of two languages (Lei, Akama, & Murphy, 2014). Therefore, the term *code switching* will be understood as an intra-sentence use of two different languages (for reviews see Heredia & Altarriba, 2001). *Language alternation* will be used to describe general between-sentence use of two languages, or conversation, and *language switching* will be used to refer to an experimental approach, or task.

In language switching tasks, the cognitive processes involved during bilingual language processing are commonly investigated by examining the difference in RTs between language conditions with stimuli in one language (either native language or second language) and conditions in which the language of the stimuli is switched (Mix) (Declerck, Koch, & Philipp, 2012). The performance differences are termed the *switch costs* and considered to be a marker for language control (e.g., Declerck, Koch, & Philipp, 2015, 2012; Declerck & Philipp, 2015b). Typically, poorer performance has been observed in language switching conditions than monolingual conditions (e.g., Bobb & Wodniecka, 2013; Costa & Santesteban, 2004a; Declerck, Koch, & Philipp, 2015; Philipp, Gade, & Koch, 2007; Prior & Gollan, 2013).

Because of between-language competition, bilinguals use inhibition of a non-intended language to allow selection of the desired language. Lexical selection involves suppression of the non-intended language, which interferes with the intended language. This is called the *Inhibitory Control Model* (ICM; Green, 1998). Language alternation, then, is regulated according to the language demands of a task. The amount of inhibition of a competing lexicon is proportional to the degree of activation of available lexicons, and it depends on the difference in first- and second language proficiency (Declerck, Koch & Philipp, 2015; Bultena, Dijkstra, & van Hell, 2015; Macizo, Bajo, & Paolieri, 2012).

When pilots or ATCOs want to speak in their second language, the inhibition of their dominant native language needs to be greater than would be the inhibition of the second language while speaking in the native language. Interestingly, however, *asymmetric switch costs* have been observed (e.g.; Campbell, 2005; Green, 1998; Meuter & Allport, 1999; Verhoef, Roelofs, & Chwilla, 2009), where there was evidence of larger switch costs when switching to the native language from the second than vice versa (e.g., Declerck, Koch, & Philipp, 2012; Meuter & Allport, 1999). Meuter and Allport (1999) suggested that larger switch costs may reflect having to overcome residual inhibition of the more dominant native language. The lexicon that has been inhibited needs more time to overcome its inhibition and recover its normal level of activation, meaning that it can take longer to overcome inhibition of the native language than the less dominant second language (Macizo, Bajo, & Paolieri, 2012).

Indeed, *second language proficiency* has been found to affect performance (Barshi & Farris, 2013; Cardosi, 1993; Estival, Farris, & Molesworth, 2016; Monan, 1991; Prinzo et al., 2008, 2010a, 2011). It appears that language alternation, especially when it is conducted by early bilinguals who grew up with two languages (Moradi, 2014), is done automatically. However, Costa and Santesteban (2004a) did not confirm the presence of asymmetric switch costs in highly proficient bilinguals. This may suggest that different mechanisms are available for learners and highly proficient bilinguals. Highly competent bilinguals may not need to apply inhibition of the non-intended language if they have developed a different mechanism of lexical access (Costa & Santesteban, 2004a). Moreover, symmetric switch costs have been observed when comprehending words in sentence context; switching to the native language was found to be easier than switching to the second language (Bultena, Dijkstra, & van Hell, 2015). Overall, there seems to be a lack of consensus regarding when and why switch costs occur, which can probably be attributed to different types of stimuli and modalities used in language switching studies.

It can then be assumed that the alternating of languages during radio communication will require additional language control than speaking in just one language would. Consequently, even if the response delays are not significant for practice, *language selection errors* can occur, causing unnecessary repeated transmissions. Potentially, performance accuracy consequences might be more serious than longer latencies caused by switching between languages. It was found that speakers made more language selection errors when switching from their second language to the native language than vice versa, and that these errors were more frequent after a short sequence of second language trials than after a long sequence of second language trials; that is, when switching occurs frequently (Zheng, Roelofs, & Lemhofer, 2018). Participants more often replaced words in the dominant language with words from the non-dominant language (e.g., Gollan & Goldrick, 2016; Gollan, et al., 2014).

In a typical language switching task (e.g., Costa & Santesteban, 2004a; Declerck, Koch, & Philipp, 2012; Philipp, Gade, & Koch, 2007) participants are asked to name a visually presented object (called a concept), called a 'picture-naming task'. The language in which the picture or digit should be pronounced is indicated by the presentation of a non-verbal cue before or simultaneously with the to-be-named concept, in a *cued language switching paradigm* (e.g., a colour or shape is presented to indicate which language should be used). Alternatively, in a *sequence-based paradigm*, the required language may be indicated by a pre-defined sequence that should be followed (e.g., switch language after every second trial: L1-L1-L2-L2-L1) (Declerck, Koch, & Philipp, 2015, 2012; Declerck, Philipp, & Koch, 2013). Sequences of endogenously triggered concepts or familiar sequences, such as naming weekdays or numbers, can be used instead of visually presented concepts, or a new fixed concept sequence may be learned prior to the experiment (e.g., Declerck, Koch, &

Philipp, 2015; Festman, Rodriguez-Fornells, & Münte, 2010). Alternatively, participants can choose the language in which they will name the visually presented concepts (Declerck et al., 2013), and auditory response cues can be used instead of visual ones, to indicate that the next concept can be verbalised (Declerck, Koch, & Philipp, 2015).

A sequence-based paradigm strives to measure *predictable* language switching. Studies involving this analysis indicated that predictability reduces the switch costs and therefore it was concluded that knowing both language and concept in advance (i.e., what is to be named) can resolve language interference (e.g., Declerck, Koch, & Philipp, 2015). Analogously, a standard routine air traffic communication can be considered as a reasonably predictable sequence of clearances, therefore potentially reducing the switch costs of an ATCO communicating in a bilingual system. Provided the ATC situation does not require the use of plain language, bilingualism would therefore seem to pose no adverse effects on the performance of bilingual ATCOs, because there would be minimal additional dual language processing costs.

A question might arise as to the relationship between the positive effect of preparation observed in language switching studies, and the negative effect of the expectation bias that is known to occur in aviation. Specifically, while predictability of a concept and language sequence can reduce RTs, can that expectation affect pilots' or ATCOs' perceptions of what was heard? According to Cushing (1995), the expectation of an instruction can prime a pilot to 'hear' the anticipated ATC instruction, even though a different instruction has been given. Grayson and Billings (1981) provide the following example from an ASRS report:

"Aircraft A was in a block altitude of 12,000–14,000 ft. The instructor pilot and student both thought the controller told them to turn left to a heading of 010° and descend to and maintain 10,000 ft. The controller requested aircraft A's altitude. The crew responded 10,700 ft. The controller stated the aircraft had been cleared to 12,000 ft, not 10,000 ft. There are two contributing causes for this occurrence: 99% of all clearances from that area are to descend to and maintain 10,000 ft, and as the instructor I was conditioned to descend to 10,000 by many previous flights. The controller may have said 12,000 ft but I was programmed for 10,000 ft." (Grayson & Billings, 1981, p. 48–49).

Understandably, expectation bias can occur regardless of the language used for air traffic communication. Nevertheless, it is possible that bilingual communication could increase this type of mistake in bilingual air traffic environments. When faster RTs are found in a research, they may not necessarily indicate better performance in practice. Consideration relative to performance accuracy must be made too.

Although most of the research in language switching has focused on visually presented stimuli, studies have also confirmed that when *listening* to sentences, language changes incur RT costs (Bultena, Dijkstra, & van Hell, 2015; Cheng & Howard, 2008; FitzPatrick, 2011; Ruigendijk, Zeller, & Hentschel, 2009). Declerck et al. (2015) compared switch costs obtained with auditory and visual stimuli and revealed that they were relatively larger with visual stimuli. However, the auditory stimuli consisted of non-speech sounds (e.g., saying “bird” when a chirping sound was heard), which seldom occur in aviation. Indeed, speech production studies are prominent in language switching paradigms regardless of whether the stimuli are single words, sentences, verbs or subjects (e.g., Christoffels, Firk, & Schiller, 2007; Costa & Santesteban, 2004a; Kroll, Bobb, & Wodniecka, 2006; Meuter & Allport, 1999; Philipp, Gade, & Koch, 2007; Philipp & Koch, 2011; Tarlowski, Wodniecka, & Marzecová, 2012).

The auditory stimuli in aviation, however, are seldom simple sounds but rather spoken utterances. Moreover, a combination of speech stimuli and a verbal response would require further consideration of how speech stimuli in different languages are processed when they are heard and when they are spoken. Costa and Santesteban (2004b) stated that the nature of the processes involved in speech production and perception are different. In word recognition, language identification is partially given by the input stimulus itself (e.g., it is phonologically encoded), but in word production, it is the speaker who intentionally chooses the target language (Costa & Santesteban, 2004b).

In auditory communication, a listener has no control over the input, which can change from one language to the other unexpectedly. When bilinguals are confronted with an unexpected language change, they detect a change during early stages of stimulus processing (i.e., 200 ms after stimulus onset) (Kuipers & Thierry, 2010). Listening seems to be cognitively more demanding than reading, given that listening takes place in real time with no visual text for reference (Vandergrift & Baker, 2015). The listener does not have the option of reviewing the information presented and has little control over the input (Vandergrift & Baker, 2015). Furthermore, spoken language is characterized by accent and speed, which can make understanding more difficult than it is for reading written text. Listening can also be more context sensitive; different intonation and emphasis can provide additional information (Mehrabian, 1981). However, somewhat contradicting results are shown in the research literature (e.g., Declerck et al., 2015). In particular, faster RTs were observed when processing stimuli were presented auditorily rather than visually (for reviews on auditory versus visual modality see Moreno, Federmeier, & Kutas, 2002).

In summary, studies have typically implemented visual stimuli combined with language production (e.g., word reading and picture naming), which can impose some constraints on the generalizability of these findings to language perception of auditory stimuli in aviation. Although the research on language switching is extensive, current knowledge is insufficient to draw conclusions about the difference in performance, if any, when aviation personnel operate in bilingual or monolingual air traffic environments. Recent studies on language issues in aviation have focused mainly on language proficiency and the composition of messages (Barshi & Farris, 2013; Cardosi, 1993; Monan, 1991; Prinzo et al., 2008, 2009, 2010a, 2010b), rather than on language switching. This thesis will help address the gap in this body of knowledge. To achieve this, languages suitable for the empirical analysis must first be selected.

3.6. Contrasting Chinese and English Language

In 2007, China reported having more than 14,000 pilots of whom 8,600 flew on international air routes (Xinhua News Agency, 2007). Previously, communicators in the cockpit were responsible for accepting clearances and communicating with other aircraft; however, technical advances have resulted in the number of aircraft crew being reduced, which in turn puts more demand on pilots’ English communication abilities (Xinhua News Agency, 2007).

English and Chinese are two of the most commonly used languages in the world, yet they simultaneously represent the two extreme examples of different language families—the tone and non-tone languages (e.g., Ge et al., 2015; Wang & Chen, 2013). This, along with the fact that China is one of the countries that uses bilingual air traffic, were the key reasons for choosing Chinese as participants' native language, in addition to English as their second language. However, the choice of Chinese language was also motivated by a more subtle reason. It was hoped that an explanation of the differences between these two languages might facilitate understanding, which could replace the anecdotal critiques of the English language proficiency of native-Chinese speaking pilots (e.g., Barshi & Farris, 2013; Knold, 2007; Michael Good Videos, 2013), and direct attention toward creative troubleshooting in communication.

According to Nordquist (2017), a *language family* is a set of languages derived from a common ancestor or *parent*. Languages with a significant number of common features in phonology, morphology and syntax are said to belong to the same language family (Nordquist, 2017). According to Brown and Ogilvie (as cited in Nordquist, 2017), there are approximately 250 established language families in the world, and over 6,800 distinct languages. Languages that belong to different language families have significantly different features.

Wang and Chen (2013) postulated that in language, different modes of thinking are embodied, which is consistent with Whorf's linguistic approach (cited in Pavlenko, 2014; as mentioned in section 3.3). An interesting example for illustration, is the phrase '舍得', which in English translates to 'willing' (Wang & Chen, 2013). In this Chinese phrase, two complementary actions are combined—to give and to take—which together create the unity of giving and receiving; that is, 'to take by giving' (Wang & Chen, 2013). The English translation 'willing' encompasses in essence only 'afford to lose'. Subconsciously, 'losing', by itself, has a more negative connotation than when it is understood within the context of receiving. The embedded concepts behind the words of different lexicons are interesting especially because thinking affects emotional experience, which affects the behaviour (David et al., 2014).

There are other significant differences related to language families that can make learning English a challenge for Chinese native speakers and vice versa. First, in Chinese, the same sounds pronounced with different tones can refer to different things, whereas in English, tone might convey emotional information, but indicates nothing about the meaning of the word that is pronounced (Ge et al., 2015). Second, Chinese language has relatively few syllables (approximately 400) in comparison to English, which has about 12,000 (Grasu, 2015). These two differences likely relate to other distinctive features, which will be reviewed in more detail within the context of aviation in the following sections.

3.6.1. Pronunciation and Accent

Accent is one of the most cited factors contributing to misunderstandings in aviation communications (e.g., Estival & Molesworth, 2012; EUROCONTROL, 2006; Molesworth & Estival, 2015; Tiewtrakul & Fletcher, 2010). Sometimes, understanding requests made by pilots or controllers speaking with a Chinese accent can be challenging (Michael Good Videos, 2013). The heavily accented English of some Chinese pilots can be caused by difficulties in pronouncing individual English words and intonation. Mandarin Chinese has no consonants except /n/ and /ŋ/ at the ends of syllables resulting in difficulties pronouncing many English words. Chinese speakers may leave off the consonant (e.g., *lab* may be pronounced [læ]; *hill* as [hi]) or may add a vowel after the final consonant, making the word one syllable longer (e.g., *lab* may be pronounced as [læbə] or [labu]) (Defense Language Institute, 1974; Grasu, 2015).

Chinese has no voiced stops, affricates or fricatives, which can make the pronunciation of the voiced set sound strange to NES. For example, the Chinese speaker may pronounce *bill* as [pil], *do* as [tu], and *get* as [ket] (Defense Language Institute, 1974). Moreover, *f* can often be substituted for *v* because the two sounds have the same articulatory position, and *sea* might sound like *she* because /s/ before /i/ may be pronounced as the Chinese /ś/ (Defense Language Institute, 1974). For these reasons, even a Chinese pilot with high proficiency in English can still have strong accent.

Understanding the accent of NES can be even harder for Chinese pilots. The Chinese language does not have an alphabet, but instead uses a logographic system where symbols represent words (Grasu, 2015). Some English *phonemes* do not exist in Chinese, and Chinese listeners may struggle to hear the difference between *l* and *r*, which can lead to misperception, for example, *rake* is perceived as *lake* and *rice* as *lice* (Grasu, 2015). The absence of an alphabet combined with a relatively small number of syllables in Chinese, the fast tempo of speech and the accent of a native-English speaker, make it difficult for non-native English speakers to understand the individual words of spoken English.

3.6.2. Number of Instructions in One Transmission

Another difference is the length of sentences; to express meaning, a long, structured sentence can be used in English, but a short sentence is used in Chinese (Grasu, 2015). In English it is possible to express several meanings in one sentence, whereas Chinese sentences are usually short with few modifiers to prevent confusion in meaning. This can be an important factor when giving multiple instructions within one transmission. Even though it is generally known that long transmissions can increase the number of errors (Barshi & Farris, 2013), whether the effects might be worse for Chinese speakers, regardless of their English-language proficiency, has yet to be explored.

3.6.3. Grammar Issues

Another difference between Chinese and English languages is their answers to negatively phrased questions, which are diametrically opposed (Defense Language Institute, 1974). In English, answers usually begin with a *yes* when a person *disagrees* with the statement in the question and *no* when the person *agrees* (Defense Language Institute, 1974). For example, “Don’t you need priority landing?”– “Yes, we do”, means they need priority landing, and “Don’t you have enough fuel?”– “No, we don’t”, means they do not have enough fuel. In Chinese, the answer usually does not begin with either *yes* or *no,* but if it does, then the answer is formulated in an opposite way while conveying the same meaning (Defense Language Institute, 1974). For example, “Don’t you need priority landing?”– “No, we do” (using *no* for *disagreeing*: “That’s not right, we must land soon”); “Don’t you have enough fuel?”– “Yes, we don’t” (using *yes* for *agreeing*: “That’s right, we do not have enough fuel”). Although these examples would not be used in standard phraseology, they can illustrate the structure of answers to negatively phrased questions, which may occur when a plain language is required. A misunderstanding can be easily prevented by the use of positively rephrased questions (i.e., “Do you need priority landing?”).

Additionally, when asking a question in English, the positions of subject and verb are the opposite of those used in a declarative sentence (e.g., "*You are* ready for departure" versus "*Are you* ready for departure?"). In Chinese, the position of subject and verb in questions is the same as the position in statements; the questions are conveyed by rising intonation at the end of a sentence, on the last word (e.g., "You are ready for departure" versus "You are ready for departure?") (Defense Language Institute, 1974). In English, when necessary, an auxiliary verb, 'do' or 'did', is added (e.g., "How did you do this?"). In Chinese, these auxiliary verbs do not exist, so Chinese speakers may forget to insert them into English questions ("How you do this?") and negative sentences. For example, in English, the sentence might be "They did not land", but if a Chinese speaker forgot to insert the auxiliary verb, it might sound like "They not landed." (Defense Language Institute, 1974).

Another issue that might cause confusion is that Chinese language is *tenseless*, meaning that the concept of time in Chinese is not handled through the use of different tense forms of verbs; instead, the expression of the temporal aspect relies on lexical (e.g., temporal adverbials) and discursive (e.g., narrative structure) means (Pavlenko, 2014). Learning the 17 possibilities (Pahlow, n.d.) of expressing past, present and future tense in English grammar can be a struggle, and may lead to mistakes such as "We land two hours ago."

A further issue is that Chinese language does not have the inflectional systems that exist in English (Defense Language Institute, 1974). For example, in the sentence "Pilot received two notices" native-Chines speakers may consider that the -s suffix on *notices* is redundant because it does not add any significant information to an utterance. The presence of number '*two*' already indicates that there is more than one notice. Moreover, as previously mentioned, the tense inflections on English verbs have no counterparts in Chinese. When time is important, time adverbs are added; the same applies for numbers. In Chinese, when number is not important, it is not indicated (Defense Language Institute, 1974). The grammar situation is different in English. For example, "The aircraft *flies* at the same flight level" and "The aircraft *fly* at the same flight level" portray two different situations in English. In Chinese, there is no plural form for nouns (Defense Language Institute, 1974), meaning that native-Chinese speakers can struggle to distinguish between mass and count nouns and use the correct plural form when the same rules are applied to English.

Several pairs of English words correspond to only one word in Chinese, that is, two or more words in English are rendered into the same word in Chinese (Defense Language Institute, 1974), including phrasal verbs, such as *pull up*. These were the last words uttered by the pilots of a China Northern Airlines aircraft, who wondered what ‘pull up’ meant, before they crashed (see section 2.5). A potential explanation beyond lack of English language proficiency may dwell in the differences between the two languages. Phrasal verbs are abundant in English, yet they do not exist in Chinese—or rather, they correspond to single-word verbs in Chinese (Defense Language Institute, 1974). Finally, Chinese language does not have articles (‘a’, ‘an’, and ‘the’), therefore, native-Chinese speakers may struggle to use them correctly in English (Defense Language Institute, 1974).

Chinese natives may experience difficulty in comprehending and speaking English as a result of interference from the Chinese language, such as code switching. The potential occurrence of code switching might not be solely a matter of English language proficiency (see the Tenerife accident in section 2.3; Tajima, 2004). The problem is not in learning English grammar, but in using English correctly in spontaneous speech, especially in stressful situations, when even highly proficient bilinguals tend to suffer the intrusion of their mother tongue. These languages may actually be processed differently in the brain (e.g., Ge et al., 2015).

3.7. Numerical Processing

The previous sections indicate that the processing of Chinese and English languages differs. This section will specifically focus on the processing of numbers. It has been found that the involvement of numerical information in transmissions results in communication errors much more frequently than in non-numeric transmissions (Tiewtrakul & Fletcher, 2010). There is the potential for the perception of numbers to represent a larger threat to aviation safety, and lead to more misunderstandings than communications without numerical information. Among the most frequent categories of information containing number forms that lead to an error—in non-native English speakers more frequently than in native-speakers—are altitude, heading, squawk number, frequency, route, and waypoint (Tiewtrakul & Fletcher, 2010). Clark and Tomato (2017) add that call sign confusion and flight level confusion were the most frequently reported events of number-related miscommunication to the Civil Aviation Authority in UK. What is unique about number processing, especially when alternating between languages, will be discussed in this section. However, the results of language switching studies are inconsistent, which might be attributed to various methodology, and inconsistency in material, stimuli, and procedure.

Several studies have been dedicated to the investigation of arithmetical operations in different languages (Grabner, Saalbach, & Eckstein, 2012; Spelke & Tsivkin, 2012), however, they did not compare different types of stimuli; that is, words/pictures vs. digits (Declerck, Koch, & Philipp, 2012). Therefore, only a limited explanation can be provided. Additionally, the findings of Declerck et al. (2012) appear to be in contrast with previously mentioned aviation practice; that is, the performance was found to be faster with digits in comparison to word stimuli. They compared performance on four stimulus sets; a digit stimulus set containing digits 1–9; a picture stimulus set (e.g., “car” in English vs. its equivalent in German, “auto”); a cognate control stimulus set (e.g., “man” in English vs. its equivalent in German, “mann”); and a control stimulus set with semantically-related items (e.g., “leg” in English vs. its equivalent in German, “bein”). Significantly smaller language switch costs were found in digit naming (mean RT 589 ms) than in picture naming (mean RT 901 ms), suggesting that digit naming in bilingual conditions required less language control than picture naming. This difference was minimized when the pictures represented cognates, suggesting that within-item phonological (sound) priming reduces language switch costs. Therefore, the difference was attributed to phonology (concerning the sound system of language), specifically, a large phonological overlap within digits and between languages (cognates).

Studies of mathematical operations in language switching tasks may provide insight into the role of language alternation in forming arithmetic operations and in problem solving in different language environments. Grabner, Saalbach, and Eckstein (2012) found longer RTs when doing arithmetic operations in language switching conditions than monolingual conditions. Additionally, accuracy was higher when problem solving did not require language switching. Grabner et al. (2012) assumed that language switch costs in arithmetic were due to additional numerical information processing; that is, calculation rather than mere language translation. They also found that solving multiplication problems was faster than solving subtraction problems. Spelke and Tsivkin (2001) observed longer RTs when solving exact number fact problems (e.g., exact addition: 54 + 48 = 102 or 92) when participants' language of training differed from the language of testing but found no language-related differences in RTs when participants solved approximate problems (e.g., approximate addition: 34 + 71 ≈ 110 or 80). Venkatraman et al. (2006) suggested that solving the same arithmetic problem in a different language requires additional processing. Grabner et al. (2012) associated the possible discrepancies in the findings among studies by using Dehaene's triple code model of arithmetic cognition (Dehaene & Cohen, 1997).

The *triple code model* proposes the existence of three different number codes in the brain; a language-independent magnitude representation, a verbal code associated with left-hemispheric language areas, and a visual Arabic number code. In other words, the model distinguishes between language-dependant exact arithmetic processing and language-independent approximate magnitude processing. This idea was supported by empirical data (Venkatraman et al., 2006), and also by studying various languages, such as that of the Pirahã tribe in Amazon, Brazil (see Venkatraman et al., 2006 for a review).

The Pirahã are an indigenous tribe of around 700 people, whose language does not have numbers (Pavlenko, 2014; Venkatraman, et al., 2006). They recognize one (*hói*, a "small size or amount"), two (*hoí*, a "somewhat larger size or amount"), and many (*baágiso,* "a bunch" or "cause to come together") (Pavlenko, 2014). They know no other quantification words (neither cardinal nor ordinal numbers), nor the concept of counting. They were not able to be taught simple counting, primarily because they lacked interest. Although they were unable to perform exact calculation, they were able of approximate number processing. This may indicate that the lack of a linguistically encoded count sequence prevents the acquisition of exact number processing (Venkatraman et al., 2006). It was assumed this supported the notion that magnitude representation is independent of language (Gordon, 2004), and that language plays a crucial role in certain aspects of calculation but not others (Venkatraman et al., 2006).

The aforementioned studies analysed number processing in languages other than Chinese (e.g., German vs. Italian, in Grabner et al., 2012; German vs. English, in Declerck, Koch, & Philipp, 2012; Russian vs. English, in Spelke & Tsivkin, 2001). Additional consideration, therefore, needs to be made regarding the comparison of numbers in Chinese and English. Pavlenko (2014) summarised several findings and indicated that the most transparent reflection of the decimal structure is found in Asian languages with roots in ancient Chinese, such as Mandarin, Japanese, and Korean.

Chinese number names are generated from digits 0–9 by the place of value using additional and multiplicative principles (Ifrah, 2000). For example, 11 is encoded as ‘ten-one’ (十一) and 20 as ‘two-ten’ (二十) in Chinese, as opposed to ‘eleven’ and ‘twenty’ in English. Shorter digit word length and shorter articulation time may also provide speakers with advantages on the digit span task (Pavlenko, 2014). Miller et al. (Miller & Stigler, 1987; Miller et al., 1995; Miller et al., 2000) found that Chinese-speaking children between the ages of four and six outperform English-speaking children on abstract counting and on counting sets of objects varying in size and arrangement. Similarly, Chinese speakers remembered more digits than speakers of English, Finish, Greek, Spanish and Swedish (Chincotta & Underwood, 1997; Stigler, Lee, & Stevenson, 1986). Pavlenko (2014) suggested that because of the different encoding processes among the languages, some languages may be more advantageous than the others for the processing of numbers.

Pavlenko (2014) also summarised findings indicating that bilinguals, regardless of their current language dominance and the environment in which they live, still use their native language for simple arithmetic operations; they solve mathematical problems faster in their native language and display native language advantage in digit reading and digit span memory. However, the length of residence in the second language context likely correlates with bilinguals’ preferred language for mental computations (Tamamaki, 1993; Vaid & Menon, 2000). When numbers are constructed more easily in Chinese than in English, the implications for native Chinese speakers performing arithmetical operations with Arabic numbers shall be considered.

Campbell’s (2005) analysis of digit naming and simple arithmetic (e.g., 2 + 2; 9 × 9) by Chinese–English bilinguals demonstrated that asymmetrical language switch costs vary with stimulus format (Arabic or Mandarin numerals), and that the asymmetry is observed both with direct naming of the digit (e.g., “8”) and indirect answering of the problem (e.g., “2 + 6”). Although the effects of language switching did not differ between additions and multiplications, the time required to produce Chinese number names was slightly slower with Arabic (731 ms) than Chinese stimuli (712 ms), whereas the time required to produce

English number names was faster given Arabic (709 ms) than Chinese stimuli (774 ms). This might indicate that if Chinese numerals automatically activate Chinese language processing, then participants would have to actively switch from Chinese to English in order to respond in English. This suggests that when communicating in English, the Arabic format of numbers may be easier for Chinese pilots to process than using Chinese characters.

Campbell's findings (2005) also revealed differences in response accuracy, whereby language errors were more frequent with Chinese stimuli (11.1%) than Arabic stimuli (9.3%), and more frequent when Chinese was cued (11.7%) than when English was cued (8.8%). Furthermore, errors were more common on switch trials (13.0% perseveration errors) than no-switch trials (5.1% switch errors) with Arabic stimuli, and it was more difficult to switch from English to Chinese (20.4% perseveration errors) than from Chinese to English (4.7% perseverations). With Chinese stimuli, it was only slightly more difficult to switch from English to Chinese (15.6% perseveration errors) than to switch from Chinese to English (11.1% perseverations).

For standard aviation communications in operations using numbers, the monolingual English air traffic environment may appear to facilitate faster and more accurate responses made by native Chinese-speaking pilots.

Part II: Methodological Considerations

3.8. Methodological Considerations

The previous chapters raised several methodological questions, which will be addressed in in this part of the literature review. General considerations will be discussed first.

In aviation, studies using flight simulator are generally preferred over computer-based experiments as they allow analysis of more realistic representations of real-life operations. However, when attempting to study the ways bilinguals perceive and process speech in bilingual air traffic environment, the focus should resort to computer-based experimental methods, because of the absence of direct, behavioural observations of the cognitive processes of spoken language comprehension (Zwitserlood, 1998). When attempting to analyse dual language processing, the understanding of performance is derived from cognitive data of language processing.

The aim of this thesis was to develop a research methodology that would simplify complex situations under the assumption that the experimental tasks would reflect the underlying cognitive operations involved in those situations. By doing so, it would be possible to isolate some irrelevant variables and focus on particular causal effects. Although a computer-based experiment can provide this control over variables, it may be developed in isolation of the context and the real-life issues that can affect the outcome behaviour. Therefore, to provide the logic whereby inferences concerning SA in bilingual verse monolingual conditions will be drawn from patterns of experimental performance, the experimental analysis will be reviewed.

Drawing conclusions from applied cognitive studies to aviation practice is a two-step process in which practice and experimental research complement each other. To affect performance in such a way that it improves safety, it must be possible to predict behaviour, and to predict behaviour, its principles must be understood (Lehto & Landry, 2013). To understand how performance changes in different language conditions entails the study of fundamental cognitive processes. On the assumption that the cognitive mechanisms are shared by the population of individuals, the inferences about the cognitive mechanisms can, in principle, be generalized beyond the sample and context under consideration (Lehto &

Landry, 2013). Making safety implications based on the knowledge of human performance capabilities and limitations, such as information processing, is not uncommon in aviation, because if these human performance capabilities and limitations are exceeded too much or “repeated enough, an accident will result” (Zellar, cited in Orlady & Orlady, 1999, p. 177).

To prevent overgeneralization of the findings, well-defined methodology analogous to an aviation context is required, so that the cognitive processes studied by the experimental task correspond to those involved in real-life activities as much as possible. Therefore, as far as possible, an experimental research approximating the reality was developed. Approximating the reality by means of a flight simulator would, unfortunately, greatly exceed the resources available for this thesis. Moreover, New Zealand as an English-speaking country does not serve bilingual air traffic, which limits the availability of bilingual pilots and ATCOs.

In principle, the research problem and the associated research questions could be investigated within the framework of either positivist or interpretivist paradigms using either quantitative or qualitative methodology (e.g., questionnaires and experiments versus case studies). Because the thesis seeks to objectively investigate whether bilinguals perform faster and more accurately in a monolingual or bilingual air traffic environment, this research is more appropriately aligned with the methodological assumptions of a positivist paradigm and quantitative methodology. However, as the aim of Study 1 was essentially to establish the research context and therefore generate ideas for the quantitative research methodology, it was essential to study the language issues in aviation within its real-life context holistically to provide a description of the current situation. For this purpose, the studies were guided by the assumptions of interpretivism, thus using the qualitative method. A mixed-method approach was adopted, allowing for areas of focus to be investigated using the most suitable method. By doing so, the opportunity to utilise methods that had not previously been employed for research in this particular field emerged.

To conclude, the following *issues* of the experimental research methodology were identified and will be addressed throughout the experimental part of this thesis:

i. To explore the impact of language conditions on SA using a computer-based experiment, the three cognitive processes that underlie SA—recognition, comprehension, and prediction—must be individually defined.

ii. Although the three cognitive processes will be analysed individually, a common procedure across the experiments is required to enable comparison of the nature and differences of these cognitive process.

iii. Given that acoustic speech stimuli will be used across the experimental tasks, the background noise must also be thoroughly considered during stimuli development, because radio communication in aviation is seldom without background noise.

iv. To compare performance between different language conditions, analyses of a listeners' response times on speech signals will require considerations of challenges related to different latencies of spoken stimuli.

3.9. The Reason for Including Signal Detection Theory

Given the issues of the experimental research methodology proposed in the previous section, the Signal Detection Theory (SDT) was chosen because it explains individual differences in decision making, and why some messages are noticed while others are not (Green & Swets, 1966). SDT can be crucial for developing and maintaining SA. Therefore, it can provide a substantial means for analysing the causes of SA errors in bilingual and monolingual air traffic environments.

SDT provides an effective method for effectively differentiating the detectability of information in different language conditions (monolingual verse bilingual) from response proclivity of an individual (Gelfand, 1998). Thus, it can help to direct the implications accordingly, whether changes need to be made on the side of language conditions (e.g., to facilitate message detectability), or on the side of an operator (e.g., response bias and training). In this way, it can facilitate decision making about the language conditions most favourable for safe operations. Moreover, SDT can represent effective overlap of the underlying cognitive processes of SA (recognition, comprehension and prediction) and the language switching paradigm.

Finally, SDT also provides a useful quantitative measure of human performance, given its background in mathematics and statistics. Its mathematics background is the primary advantage of SDT because it allows objective empirical analysis. However, because the extent of the theory's mathematical foundation is beyond the scope of the thesis specialisation, discussion of the mathematical aspects of SDT in the following sections will be limited to simplified explanations of the most relevant aspects.

3.10. SDT, SA, and Bilingualism

The aim of most ground-to-air communication is to provide pilots and ATCOs with necessary information to be able to achieve and maintain SA. The information can be obtained in two language conditions, monolingual English and bilingual. Utilizing SDT can provide empirical evidence as to which of these two language conditions facilitates faster and more accurate responses, and, arguably, better SA. According to Edgar et al. (2018), SDT has been previously used to provide a performance-based measure of SA. For example, Allendoerfer et al. (2008) investigated ATCOs' decision making about the presence of potentially hazardous situations during monitoring of the traffic situation in their sectors and simultaneous monitoring of the output of an alerting system. Ability to detect information is affected by the intensity of the information (e.g., message intelligibility on a noisy background) and operators' ability and their physical and psychological state or predisposition (e.g., how alert they are).

Presumably, the use of two languages for air traffic communications may decrease the recognition of information, given the larger cognitive demands of alternating between two languages (language switch costs). Moreover, some information will be relevant, and some will be irrelevant, or distracting. Indeed, the use of a language that a crew does not understand will arguably represent noise to them, given the language barrier. Homophony can also impair detection of information. Many perceptual errors in aviation communication can be attributed to the ambiguity of the information presented (McNicol, 1972). Phrases such as 'I'll let you know' vs. 'Let him go', or 'Last of the power' vs. 'Blast of power', 'Two' vs. 'To', and 'On the hold' vs. 'On the go' are just a few examples of such real-life instances (Cushing, 1994).

Language and ambiguity are just two examples of *external noise* related to the understanding of communication. There are many other possible sources of external noise related to headphones, quality of transmissions, static interference, noise from an aircraft engine, and other circumstances. Non-native English listeners appear to be more affected by a noisy environment than native-English listeners (Mattys et al., 2012). In the absence of noise, auditory word discrimination was found to be almost equivalent for both native and non-native English speakers; with an increased level of noise the performance deteriorated in both groups, however, the effect was worse for non-native English speakers (Gat & Keith, 1978).

Using microphones for radio communication can clip part of a message; when keying the microphone, the first syllable from a call sign may be clipped. Consequently, the call sign might be misunderstood, causing another aircraft to execute the instruction. For example, an aircraft with the call sign 'Echo Alpha Kilo' may accept a message meant for 'Tango Alpha Kilo' (McMillan, 1998). In a worse-case scenario, a flight crew missed the first digit of a flight level instruction, so that instead of hearing 'one six thousand' (16,000 feet), they heard 'six thousand' (6,000 feet) (Cushing, 1994). This event ended safely—a pilot from another aircraft detected the incorrect read-back and addressed it. That pilot was situationally aware.

While external noise can be reduced, for example, by using noise cancelling headphones (e.g., EUROCONTROL, 2006; Terenzi, 2006; Simpson, et al., 2005), there is little or nothing that pilots can do to reduce internal noise. *Internal noise* refers to the neural activity in the brain that determines a pilot's impression about whether information is present or not (Heeger, 2003). This internal noise is inherent; that is, even without any stimulus there will be some internal noise in the pilot's or ATCO's sensory system (Heeger, 2003). McNicol (1972) provided an example whereby a person can see twinkling spots of light when sitting in a dark room. Neurons in the central nervous system can fire spontaneously without external stimulation (McNicol, 1972). As a result, even when there is no actual stimulus (e.g., a message) present, if a pilot expects one, the brain can create random neurological excitations from very weak internal stimuli, leading to the familiar situation of '*hearing what they wanted or expected to hear*'. The concept of internal noise carries with it the implication that all our choices are based on evidence, which is to some extent unreliable (or noisy) (McNicol, 1972).

The above discussion of the overlap in the information and noise implies that decisions will always be made with some degree of uncertainty, which suggests that performance will always be open to perception errors, even in very experienced aviation professionals. Four possible outcomes can be observed, based on the combination of the presentation of information and the response made to it (see Table 1). For example, for a communication to be successful, pilots must recognize that a transmission was for their aircraft, comprehend it and execute the instructions correctly. Pilots and ATCOs can initiate appropriate *responses* only after successful recognition of a particular message (*information*) transmitted amongst many others. As can be seen in Table 1, the four possible reactions are: *hit* (the information is relevant for a particular aircraft and the pilot responds), *miss* (information was relevant for an aircraft but its crew did not respond), *false alarm* (pilots executed an instruction issued to another aircraft), and *correct rejection* (relevant information was absent and pilot did not respond).

Table 1

Information–Response Matrix

		Response	
		Yes	No
Information	Present	Hit	Miss
	Absent	False Alarm	Correct Rejection

Consequences of various responses can range from false alarms costing money and time to missing the important information costing human lives (Baddeley, 1997; MacDonald & Balakrishnan, 2005; Swets, 2001). Hits and correct rejections are desirable, and are likely to be consistent with SA; whereas false alarms and misses can indicate a lack of SA, and as such, pose a threat to aviation safety. However, as McNicol (1972) stressed that false alarms can reveal as much about decision processes as correct detections, the analysis of false alarms and misses can also help us understand how SA is achieved.

3.11. Yes–No Procedure

The previous section indicated that the four possible outcomes of response decisions made about information (ATC message) can best illustrate the favourability of either language condition in terms of accuracy and, thus, operational safety. To obtain these data, a Yes–No detection task was utilised.

In a Yes–No detection task, a listener is asked to respond by selecting one of the two permissible response alternatives (Green & Swets, 1988), either *Yes* or *No*. They are not allowed to make any other responses, such as "I don't know", or to skip to another stimulus. Importantly, the response is made after hearing each of the stimuli. In other words, after the presentation of an auditory stimulus, the participant's response follows. This 'stimulus presentation, participant's response' sequence is the same throughout the task. An optional phase of feedback provided to a participant—informing them whether the response was correct—can be inserted in between the participant's response and the presentation of the next stimulus. This phase has meaning and purpose during practice trials, allowing participants to fully comprehend the task.

The language switching paradigm—that is, the presentation of acoustic speech stimuli in monolingual and bilingual conditions—can be investigated using a Yes–No detection task. Specifically, participants would make a response after hearing every single word stimulus, whether in a monolingual or bilingual; that is, language switching condition. RTs, reflecting the speed of performance, and the four response outcomes, reflecting the accuracy of performance, are recorded. Together, they can reflect the impact of language alternation on performance.

To sum up, utilising the Yes–No detection task in a language switching paradigm provides the opportunity to measure the following *indices of performance*: RT, type and percentage of errors, sensitivity and response bias, and language switch costs. These and their related challenges will be described in more detail in the following sections.

3.12. Response Time: The Challenge

According to Woodworth (as cited in Green & Swets, 1966), RT is one of the most suitable variables for experiments focused on performance. It is a useful measure in two ways (Woodworth, cited in Green & Swets, 1966, p. 326): "as an *index of achievement* and also as an *index of the complexity* of the inner process by which a result is accomplished, for the more complicated the process, the longer time it will take." RTs vary as a function of the complicacy of the task (Green & Swets, 1966).

There appears to be a challenge associated with the way RTs are measured when used as an index of achievement. Two possibilities exist for measuring the RT, depending on when the response is executed. The response can be carried out either after the whole stimulus was heard or while the stimulus was playing. The former refers to the *off-line method* of measuring the RT and the latter to *on-line method* (Marslen-Wilson, 1985).

The on-line method reflects the immediate properties of the analysis processes of the speech input (Marslen-Wilson, 1985), whereas in off-line analysis, the post-perceptual processes are predominantly analysed. Importantly, the on-line analysis denotes the "minimum time-window over which the process of speech analysis can operate" (Marslen-Wilson, 1985, p. 56). Therefore, in discovering processes underlying language control, it is preferable to use the on-line method of measuring RT, as it allows the exploration of real-life speech recognition performance in the monolingual condition, and comparison with the bilingual condition. The on-line method makes the possible responses to the stimuli unrestricted by setting the *timing of response onset and offset*.

Several studies using the on-line method (e.g., Holcomb & Neville, 1991; Marslen-Wilson, 1985) have indicated that processing of words started before the spoken word had been heard completely. Additionally, studies using Chinese and English word stimuli have shown that word (Pavlenko, 2014) and sentence length (Venkatraman, et al., 2006) were longer in English than in Chinese. Because of the different length of Chinese and English stimuli and their articulation times, it is unclear which mechanism actually causes the switch costs. In other words, the switch costs can be attributed to the difference in the stimuli length, to the language processing of these languages, or to a combination of both. Because of this ambiguity, if the method failed to distinguish potential influence of different durations of the stimuli, it could fail to recognize the true direction or cause of switch costs.

The differential influence of stimuli length on RT and language processing does not appear to have been examined in the available literature. It could be argued that if the different length of stimuli in different languages will be constant throughout the studies, it is unlikely to cause any methodological problems. However, it does not necessarily mean that the obtained result will be correct but rather, that measurement error is constant over the measurements. Shorter Chinese spoken words than English words can result in the observed

RTs always being longer on English words. Yet, it does not automatically reflect the cognitive processing time, which, presumably, should be independent from the duration of the stimuli. Moreover, any influence of stimuli length could vary depending on language conditions.

In principle, this assumed potential measurement error can be corrected, and thus prevent bias in results, by simply subtracting the duration of the stimulus from the corresponding RT. This would be a new way of RT measurement, and as such, it does not exclude the possibility of limitations. Nevertheless, this approach could be crucial, because the obtained findings would indicate which language condition facilitates faster RTs when isolated from the influence of different durations of stimuli articulation. Thus, the choice of the RT measure can have considerable consequences for the ultimate interpretation of the findings.

The understanding of RT as an index of the complexity of inner processing by which a result is accomplished, may have implications for the measurement of SA. Donders (cited in Sternberg, 1969) proposed that the RT reflects the speed of mental processes, making it possible to differentiate different stages of information processing, and thus different levels of SA. Donders (cited in Sternberg, 1969) introduced the subtraction method for analysing the RT into its components, and thereby created the opportunity to analyse the corresponding stages of processing. To use the subtraction method, two different tasks are constructed, where the second task is thought to require all the mental operations of the first, plus an additional inserted operation (Donders, cited in Sternberg, 1969). The difference between the mean RTs of the two tasks is interpreted as an estimate of the duration of the inserted mental operation. Theoretically, the RT can be decomposed to individual elements.

In this thesis, which employs a cognitive approach towards SA, the idea of decomposition of RT might be especially significant. SA consists of three consecutive levels—recognition (level 1), comprehension (level 2) and prediction (level 3)—with the higher levels including the previous, less complicated level of processing. A pilot must first recognize a stimulus and comprehend it before it can be used for making predictions about the future state. Hypothetically, then, the prediction RT could be decomposed to recognition RT and comprehension RT, whereby the contribution of these two processes of recognition and comprehension, in controlled experimental settings, would be isolated.

The decomposition of RT could potentially allow a more precise comparison of the three cognitive processes involved in the achievement of SA. Although decomposing SA into individual levels would make no contribution to the knowledge of SA in a real-life situation, it might be useful to explore the extent to which the prediction RT can be explained by the other two processes. However, the method of RT decomposition can raise concerns. Therefore, this approach can be understood as a complementary perspective, which can raise new research questions and inspire further discussion and research, but it does not aim to deny the conventions in RT measurement.

In fact, these two ways of measuring RT—the subtraction method and the decomposition method—are credible only when both the presented assumptions of the effect of word length in different languages and the complexity of the cognitive processes are valid. Even though this kind of approach to RT measurement has not been used before, the potential limitations were considered reasonable when the method of RT measurement will be applied in accordance with the following two *criteria*: (i) the RT measurement must be consistent across all experiments; and, (ii) the RT measurement method will be used experimentally and any conclusions will be tentative.

3.13. Markers for Language Control

Switching between languages is associated with processing time costs (e.g., Bultena, Dijkstra, & van Hell, 2015). Generally, the literature distinguishes between two markers of language control; switch costs and mixing costs (e.g., Christoffels, Firk, & Schiller, 2007; Declerck et al., 2015; Declerck & Philipp, 2015; Declerck et al., 2013; Los, 1999, 1996; Philipp et al., 2008; Verhoef, Roelofs, & Chwilla, 2009). However, the definitions of each marker and their measurement is unclear in the literature. For example, Declerck et al. (2015, p. 378) defined mixing costs as “contrasting the performance between pure language block, in which only one language is relevant, and mixed language block, in which participants switch between two (or more) relevant languages.” This is consistent with Los (1999, 1996), but contrary to Christoffels et al. (2007, p. 193), who defined “the difference in naming latencies between switch and non-switch trials” as switch costs. However, switch costs according to Declerck and Philipp (2015, p. 167) refer to the decrease in performance “when two consecutive trials require production in a different language”; that is, when two monolingual trials in different languages follow each other. To avoid confusion, for this study, the two markers will be defined as follows:

The *mixing costs* will be obtained by contrasting the performance between the two languages that are randomly alternated in the Mix condition. The *switch costs* will be obtained by contrasting the performance between either of the monolingual conditions (either participants' native language in the L1 condition, or second language in the L2 condition) with the Mix condition. As such, the mixing costs refer to the difference in the processing demands of the two languages within the bilingual condition (reflecting language control in the Mix condition), and the switch costs refer to the performance speed difference between monolingual and bilingual conditions (reflecting language control between these conditions).

Each of the language conditions (whether monolingual L1 or L2, or the Mix) consists of a *sequence of stimuli*, which are acoustically presented to participants. In the L1 condition with Chinese (participants' native language) spoken stimuli, the stimuli follow each other in a sequence that can be expressed as *L1→L1*. In the L2 condition with English (participants' second language) spoken stimuli, the stimuli follow each other in a sequence that can be expressed as *L2→L2*. In the Mix condition, with alternating Chinese and English word stimuli, a stimulus in Chinese may follow a stimulus in English (*L2→L1*), or vice versa (*L1→L2*).

To determine whether it is faster to process native language stimuli in a monolingual or bilingual condition, the switch costs are calculated using the following procedure. To calculate the switch costs for Chinese (native language) stimuli, the mean RT measured in seconds on stimuli in the L1 condition ($mean\ RT_{L1\rightarrow L1}$) is subtracted from the mean RT on native language stimuli in the Mix condition ($mean\ RT_{L2\rightarrow L1}$). A similar formula applies for calculating the switch costs for processing the second language stimuli (see Figure 1).

$$Switch\ costs\ for\ L1 = mean\ RT_{L2\rightarrow L1}–mean\ RT_{L1\rightarrow L1}$$

$$Switch\ costs\ for\ L2 = mean\ RT_{L1\rightarrow L2}–mean\ RT_{L2\rightarrow L2}$$

Figure 1. Formulas for calculating the switch costs.

The mixing costs refer to the difference in RTs on native Chinese language (*mean* $RT_{L2\rightarrow L1}$) and on second language (English) stimuli (*mean* $RT_{L1\rightarrow L2}$) within the Mix condition. The mixing costs may determine whether it is harder to switch from the native language to the second language in bilingual air traffic conditions, or whether it is harder to switch from the second language to the native language. Slower RT when switching to native language stimuli will indicate asymmetric mixing costs, what means having to overcome residual inhibition of the more dominant native language. The mixing costs will be calculated using the formula in Figure 2.

$$\textit{Mixing costs} = \textit{mean } RT_{L1\rightarrow L2} – \textit{mean } RT_{L2\rightarrow L1}$$

Figure 2. Formula for calculating the mixing costs.

3.14. Discriminability Index and Decision Criterion

The concept of SDT distinguishes between the effects of the characteristics of a stimulus (e.g., its loudness, frequency and ambiguity) on performance and the effects related to an observer (e.g., inter-individual differences between pilots and between ATCOs). The characteristics of a stimulus refer to the degree to which the information is discriminable from noise (Heeger, 2003), which is also sometimes described in the literature as the sensitivity of an operator to distinguishing information from noise (Stanislaw & Todorov, 1999). Inter-individual differences refer to the decision criterion, which depends upon the benefits and costs of the various decision outcomes, especially the consequences of false alarms and misses (Lynn & Barrett, 2014; Swets, Tanner, & Birdsall, 1961). The variation in performance speed and accuracy is closely related to and dependent upon the variation in the decision criterion adopted by an operator for making a response and upon the discriminability of a stimulus from noise.

The discriminability of the stimuli, expressed by the *discriminability index* (d'), is an estimate of the strength and clarity of the information (Heeger, 2003). Its value does not depend upon the criterion the operator adopts but is based merely on the physical characteristics of information and noise (Heeger, 2003). However, when the task is easy, the information is also well separated from noise. Generally, discriminability should decrease as the physical similarity between the stimuli increases. For example, when the noise in a cockpit is very high then the discriminability of a radio call will decrease. Similarly, the basic concept of Weber’s law is that increasing the volume of a radio will have a positive effect of increasing hit and correct rejection rates (Green & Swets, 1966); that is, the higher the intensity of the background noise, the higher the intensity of the speech would have to be in order to maintain its understandability. However, intensity can be increased only up until a certain threshold, at which point speech becomes incomprehensible and the overall volume physically uncomfortable.

Even when the discriminability of information is high, two pilots in the same situation and with similar flying experience may respond differently, given their individual differences in adopting the *decision criterion* (C). The aim is to find a *decision criterion* that will lead to the most optimal decision possible (given the nature of the situation). Training and experience can positively affect an operator’s decision making. In response to a situation, the operator decides (based on their decision criterion) whether the information was presented (*yes* response) or not (*no* response). Depending on situation and experience, the criterion can be shifted more towards a *yes* or a *no* response; it does not remain unchanged. It is desirable to adjust the choice of the criterion according to the situation. Therefore, *decision bias*, or the tendency to respond *yes* or *no*, does not necessarily represent a negative feature. The following two examples can illustrate this.

First, in the case where a signal (or information) is very rare, the optimal decision criterion would be towards a *yes* response, so that when it occurs, the pilot always responds. This can be the case of automation. When automation rarely fails, and most of the time functions very well, pilots tend to trust its signals. When a signal that indicates the presence of a hazard is issued, pilots tend to consider that signal to be reliable (*yes* response), and not noise. This tendency can increase the number of hits but also the number of false alarms, potentially leading to preoccupation by automation. This was arguably the case in the Eastern Air Lines Flight 401 accident, where the flight crew was so preoccupied with a burnt-out landing gear indicator light that they failed to detect that the aircraft was gradually losing altitude, and crashed (NTSB, 1973). Alternatively, when an automation system does not issue any signal, pilots do not question the lack of a signal. Dzindolet et al. (2003) found that participants initially considered automation as trustworthy and reliable. After observing automation failure, participants then distrusted even reliable automated technology, unless an explanation on why the automation failed was provided. This illustrates how easily the decision criterion can be influenced and changed over time.

Second, where distracting information is very frequent, the decision criterion would be towards *no* response. Where pilots do not understand languages other than English spoken in an airspace, their tendency toward *no* response can increase, meaning that they pay less attention to messages they do not understand. Consequently, the number of misses of an ATC message directed to their aircraft may increase.

According to MacDonald and Balakrishnan (2005), in every individual situation there is only one choice of decision criterion that would lead to an optimal performance and all others would be suboptimal to some degree. Some pilots may choose to respond any time there is the slightest indication of a signal, thereby applying a more *liberal* criterion for their judgement. Others may apply a *stricter* criterion and decide to react only when they are very certain that it was a signal. This judgement according to the criterion determines the *response bias* of an individual. Operators who apply liberal criterion may have faster responses than operators who take longer to decide whether a signal was present. Therefore, it can be assumed that liberal criterion can, in certain circumstances, also indicate hasty, or impulsive decisions.

A person's response tendency depends on how the potential consequences of different types of errors, such as a consequence for missing a message or the penalty for a false alarm, are perceived (Heeger, 2003). A pilot may feel that missing a message may mean the difference between life and death, whereas a false alarm may result only in repeated transmissions or perhaps feeling a little embarrassed. In the case of the Zagreb accident (AAIC, 1976), a missed instruction to climb issued to Inex Adria in Serbo-Croatian language, which British Airways did not understand, is an example of a potentially fatal consequence. However, the cost of a false alarm can also potentially be high, such as when a flight crew take a clearance or instruction not intended for them.

Some pilots may feel that unnecessary repeated transmissions cause frequency congestions and confusions. Some pilots may also not want to offend or irritate ATCOs, especially if the ATCOs are known for being rude or aggressive (Gladwell, 2008). Finally, some pilots may feel that a message, if there really is one, will be repeated by the ATCO if they do not receive a read-back. Monan’s (1988) report on hear-back problems pointed out that pilots’ implicit expectation is that if they say something wrong, an ATCO will correct them, which then leads to reduced performance in pilots’ own active listening. Additionally, the lack of a response from an ATCO is often considered silent confirmation that the read-back was correct (Monan, 1988). Both pilot and ATCO must make yes–no decisions about the correctness of the reception of the message.

Stress causes the shift of criterion towards a *yes* response, thus increasing the number of correctly identified signals, but also the frequency of false alarms (Daniel, 1984). The lowest probability of signal detectability was found immediately after a signal presentation, because, then, participants probably least expected another signal (Daniel & Pikala, 1976).

The decision criterion can also be affected externally using different instructions (Green & Swets, 1966). For example, the test instruction for applying the strict criterion may be “respond as accurately as you can”, to maximize the percentage of correct responses. For the liberal criterion, the instruction would be “respond as fast as you can.” The instruction can amplify the balance between the value of performance speed and the value of correct responses. Moreover, when the instruction would emphasise the speed of the responses, the on-line method of measuring the RT should be applied, so that participants are able respond as soon as they decide about their answer. This is in accordance with the RT measurement discussed in section 3.12.

The goal of SDT is to estimate the value of the two parameters, discriminability and criterion, from the experimental data and create a model of participants’ responses that maximises the number of correct responses and minimises losses (McNicol, 1972). Asking for repetition of the message is an effective strategy for acquiring necessary information to increase the likelihood of getting either a “hit” or a “correct rejection.” The number of repetitions required to reach a correct response (especially in a noisy environment) can also be a measure of communication performance in bilingual, compared with monolingual, air traffic environments.

3.15. Background Noise and Speech to Noise Ratio

General issues regarding the noise, and its effect on detection performance, were discussed in section 3.10, whereas its challenges, especially for the subsequent development of the methodology, are discussed in this section.

Listening in a cockpit environment always includes coping with some background noise, and it may not always be completely clear what was said. Noise levels in most commercial aircraft cockpits range from 85 to 100dB SPL (sound pressure level) under normal operating conditions (Ericson & McKinley, 2001), which can make intelligibility and comprehension of transmissions hard. The degree of discriminability of a signal depends upon the *speech to noise ratio* (SNR) (Swets, Tanner, & Birdsall, 1961). SNR is the measurement of the audio signal level compared to the noise level present in the signal and is typically measured in dB (Chan & Simpson, 1990). For example, a SNR of 10dB means that the level of the radio signal is 10dB higher than the level of the noise. For signals to be clearly perceived, their intensity should exceed the intensity of noise by approximately 6dB (Lomov et al., 1983). According to Chan and Simpson (1990), pilots prefer to set intercom listening levels to a SNR of 0–10dB.

A challenge of SNR is that if the background noise changes over time it could affect performance, which could itself change as a function of the changes in background noise. The solution to this undesirable effect may involve setting up experimental conditions with a fixed SNR, thus creating an equal set of acoustic stimuli.

Part III: Summary

3.16. Summary of the Literature Review

Since the establishment of bilingual air traffic communications, the need to improve English language proficiency has been identified. The reasons were rooted in the aircraft accident investigations and outcomes, where language and bilingualism were considered contributing factors. It was recognized that pilots who do not understand non-English language communications in their airspace may be left out of the communication loop and be unaware of the air traffic in their vicinity. This might have affect the decision to establish monolingual English language operations at some international airports. However, the performance of bilinguals when they alternate between two languages has not been investigated, except in simulation studies (Borins, 1983), which revealed some impairments related to language switching.

Language alternation requires processing of two languages, and therefore, language control which will ensure that bilinguals communicate in an intended language. Previous sections indicated that investigation of language processing calls for an experimental approach, given the lack of behavioural observations of the language control involved in spoken language comprehension (Zwitserlood, 1998). Therefore, the language switching paradigms were reviewed. Despite language switching being a well-researched area, the findings do not provide sufficient insight into language alternation in aviation. This review indicated a number of gaps relating to the stimuli used, how they were presented, and how the participants responded.

The transmissions in aviation are seldom pictures, single words, non-words, or even the simple sounds that are typically used in language switching studies; rather, they are sentences with a defined structure. Similarly, the responses made by pilots in aviation are beyond simply naming what was seen (e.g., as required in the picture naming paradigm). Moreover, when considering the concept of SA from an operator-focused perspective (see section 3.2), greater importance in investigations might be given to the analysis of language perception (i.e., information acquisition), rather than language production. This indicates a need for a non-verbal response to speech stimuli, as was explained in section 3.5. Therefore, to make suggestions for the improvement of communication in bilingual aviation environment from the language switching findings and observations, further investigations are required. The current research study was developed in response to this need to investigate the SA of bilinguals in a language switching task.

The reviewed literature indicates the importance and benefit of a multidisciplinary approach in tackling the problem of language alternation in aviation. To test the three cognitive processes underlying SA—which is an operator-focused approach to bilinguals' SA measurement (see section 3.2)—the SDT in language switching paradigm were used. Exploring language switching as the primary task can provide fundamental understanding of how the bilingual environment affects the performance of bilinguals. Three yes–no signal detection tasks were developed to test recognition, comprehension and prediction. They were developed using a language switching paradigm implementing auditory speech stimuli presented in an unpredictable language sequence combined with a non-verbal response made by participants using a keypress, which can be analogous to microphone keying.

Use of the language switching paradigm to compare bilinguals' recognition, comprehension and prediction in monolingual conditions with bilingual conditions led to the need to consider the form of RT measurement (see Section 3.12). On an empirical level, the obtained findings would inform us about the effects of language switching on the three cognitive processes. On a practical level, the findings will help specify characteristics of the language alternation of bilinguals during bilingual language control.

The thesis used a combination of several approaches to obtain evidence answering the proposed research questions with the aim of improving aviation safety related to bilingualism. As the initiative of establishing a single aviation language remains challenging (ICAO, 2010), the main contribution of this thesis is to provide empirical evidence of the concrete effects of language alternation on the performance of bilingual aviation personnel.

3.17. Identification of the Research Problem

Chapters 1, 2 and 3 point to the existence of a research problem, that *the use of more than one language in an airspace may adversely affect performance of bilinguals*. However, there appeared to be lack of empirical knowledge as to what these effects might be. This study, therefore, aims to investigate the effects of different language conditions on the recognition, comprehension and prediction performance of bilinguals. By providing a set of objective measures (obtained using SDT) for evaluating the efficiency of bilingual conditions, this thesis also aims to help to bridge the existing gap in the empirical investigation of bilinguals' performances.

3.18. Identification of the Research Questions

Critical consideration of the academic literature led to the development of the following research questions, presented according to the study in which they are addressed:

Study 1: Pilot and ATCO Current Language Experiences

(1a) What was the English language proficiency of non-native English speaking pilots and ATCOs participating in Study 1?
(1b) How frequently was bilingual air traffic radio communication experienced?
(1c) What problems related to bilingual air traffic have pilots and ATCOs experienced?
(1d) Does bilingualism affect the SA of bilinguals and monolinguals?
(1e) What are the perceived consequences, if any, in relation to operating in a bilingual air traffic environment?
(1f) Do non-native English participants use their native language while on duty, when no radio calls are made?

Study 2: Call Sign Recognition

(2a) Do speed and accuracy of call sign recognition differ between monolingual and bilingual conditions?

(2b) Does call sign similarity affect call sign recognition in monolingual and bilingual conditions?

Study 3: Error Identification

(3a) Do speed and accuracy of error identification differ between monolingual and bilingual conditions?
(3b) Is there a difference in performance speed and accuracy when responding to correct information and when identifying a mistake?

Study 4: Prediction

(4a) Do speed and accuracy of prediction differ between monolingual and bilingual conditions?
(4b) How many steps ahead can the presence of a given target event be predicted from a sequence of events, before performance speed and accuracy are affected?

Study 5: Listening to Radio Calls over Background Talk

(5) Does listening to two simultaneous messages affect performance on call sign recognition, error identification and prediction in monolingual and bilingual conditions?

Study Six: Sterile Cockpit

(6) Is there a difference in performance speed and accuracy between the tasks performed with and without background talk?

CHAPTER FOUR

Study 1: Pilot and ATCO Current Language Experiences

4.1. Introduction

"…We were told… to taxi to Runway 6... When we were exiting the ramp area.., the ground controller stated, "Just for your information there is an opposite direction Cessna...." I then stopped the aircraft… The Cessna taxied past us... At no time were we told to hold short of the runway… If I had continued and turned right onto the runway we may have had a head on collision... I cannot say if the ground controller told the Cessna to give way to us as all radio chatter to others was in French…. I believe a huge factor in this was we were the only subjects on the frequency that were speaking English or were spoken to in English. All other radio transmissions… were in French... Had the controller been speaking in English we might have known this Cessna was going to be taxiing past us way before I was forced to take evasive action… As a pilot I gather much information about my surroundings and the airport environment by listening to ATC talk to other aircraft and ground equipment…" (ASRS Report No. 995712)

The above example, which occurred in Quebec in 2012, amply illustrates how critical bilingual communication can be in aviation safety (ASRS Report No. 995712). This is just one example of how bilingualism on a radio can affect SA. However, it illustrates the experience of monolinguals and appears to lack insight into the experience of bilinguals. Additionally, although incident reports can probably provide a compelling link to current communication problems, they are limited to safety occurrences, and, as such, may omit the critical source of understanding of daily-life experiences with two languages in aviation.

Three studies (Prinzo et al., 2008, 2010a, 2010b) exploring the experience of monolingual pilots in a bilingual air traffic environment were conducted before the application of ICAO's LPRs, on March 5, 2011. A series of questions was raised, such as what has changed since the LPRs implementation due date, whether the English language proficiency of ESL aviation personnel has improved, and whether the implementation of LPRs affected the frequency of use of two languages for air traffic communications. To explore the current situation in aviation regarding the use of two languages, Study 1 was designed to investigate the recent experience of pilots and ATCOs with language alternation when operating in bilingual air traffic environments. The following research questions were formulated:

Question 1a: What was the English language proficiency of non-native English speaking pilots and ATCOs participating in Study 1?

Question 1b: How frequently was bilingual air traffic radio communication experienced?

Answers to these questions may help to identify current problems related to bilingualism, for investigation in subsequent studies. Among the findings from Prinzo et al.'s (2008, 2010a, 2010b) studies investigating the experiences of native-English speaking pilots operating in a non-native English speaking airspace, was a report that language alternation affects SA. In response to these findings, this study aimed to extend these analyses to bilinguals' experiences of the effects of language alternation on SA, as well as on the experiences of ATCOs, as well as pilots.

Question 1c: What problems related to bilingual air traffic have pilots and ATCOs experienced?

Question 1d: Does bilingualism affect the SA of bilinguals and monolinguals?

Question 1e: What are the perceived consequences, if any, in relation to operating in a bilingual air traffic environment?

The previous chapters suggested that even if radio communications were to be conducted only in English, the issues related to language alternation would be unlikely to be completely solved. Language alternation may also occur when alternating communication between radio calls and the cockpit or control room. In practice, it seems likely that pilots and ATCOs are most likely to communicate with each other in their native language when there is silence on the radio frequency, such as during long flights. To test whether this is true, the final question was:

Question 1f: Do non-native English participants use their native language, while on duty, when no radio calls are made?

4.2. Method

4.2.1. Participants

An online survey was completed by 214 aviation personnel (pilots: *n* = 181; 84.58%; ATCOs: *n* = 33; 15.42%). Participants were NES (*n* = 66; 30.84%) or non-native English speakers who had ESL (*n* = 148; 69.16%). The ESL respondents were further classified on the ICAO English language proficiency scale, where 14.19% (*n* = 21) had attained level 4 (Operational), 45.95% (*n* = 68) had attained level 5 (Extended), and 39.86% (*n* = 59) had attained level 6 (Expert) made up the remainder of non-native English speakers.

4.2.2. Materials

To collect data, an online survey was developed to provide participants high flexibility in responding and easy access beyond country borders, to obtain as many and as wide a range of responses as possible. To explore the representativeness of the sample, participant demographics were considered. Demographic information was considered to determine whether the sample included the vast majority of experience levels. For example, the perception of the effects of bilingualism can differ between ATCOs who mostly use their native language, because international flights are infrequent in their sector, and ATCOs who mostly serve international air traffic. Similarly, experiences of pilots who mostly operate in Europe may differ from those who conduct most of their flights in English-speaking countries, because exposure to a bilingual air traffic environment is decreased.

The representativeness of the sample, measured by demographic data, influenced the development of the survey questions. It was considered more relevant to ask the native language of pilots' than their origins or the base of their operation. The rationale was that it is quite common for pilots to cross multiple state borders, and therefore the country of operation may not reflect the variability of responses and representativeness of the sample. The representation of native languages within the sample might better reflect this variability, at least, in the bilingualism sense. Sample representativeness was assessed according to the variety of language families. This rationale was in accordance with Lei, Akama and Murphy (2014), who proposed several dimensions that can be considered as measures of bilingual abilities, including structural distance between the two languages (expressed by the language families), the degree of proficiency (expressed by the level of ICAO proficiency rating), and the context of acquisition and/or learning. All three categories were included in the survey.

The survey was designed using Google Forms, and met the following *criteria* for survey development. First, because different questions were asked of native-English and non-native English speakers, and pilots and ATCOs, the survey had to be tailored to specific participants, so that they could not read questions that did not apply to them or were irrelevant, such as asking an ATCO the type of flight, or asking a native-English speaker when they learnt English. To tailor the survey in this way, skip logic must be used. Skip logic is a feature based on conditional branching that changes the question that respondents see next based on how they answered the previous question (SurveyMonkey, n.d.). The skip pattern varies based on pre-defined rules. Second, the online survey must allow various types of responses, including simple yes–no checkboxes, dropdown choice menus, the ability to make several choices where applicable, and to write comments in blank spaces. The last criterion was the ability to organize responses to make analysis controllable, and to provide analysis of basic data, such as frequencies and the distribution of responses within a group.

4.2.2.1. Survey Form

Pilots and Controllers Language Experiences Survey

"This project has been evaluated by peer review and judged to be low risk. Consequently, it has not been reviewed by one of the University's Human Ethics Committees. The researcher(s) named in this document are responsible for the ethical conduct of this research.

If you have any concerns about the conduct of this research that you want to raise with someone other than the researcher(s), please contact Dr Brian Finch, Director (Research Ethics), email humanethics@massey.ac.nz"

Please choose:

○ Air Traffic Controller
○ Pilot

NEXT

Air Traffic Controllers
How many years have you been working in ATC as controller?

What part of ATC do you mainly work in?

○ Ground
○ Tower
○ Departure
○ Radar
○ Arrival
○ Approach
○ Other:

BACK NEXT

Pilots
Approximately how many flight hours have you logged (all types of aircraft together)?

What routes do you mainly fly (please choose)?

○ Domestic (not crossing international borders)
○ International short haul (crossing one international border)
○ International long haul (crossing two or more international borders)
○ Intercontinental (crossing continental border(s))

BACK NEXT

Is English your native language?

○ Yes
○ No

BACK NEXT

Native-English speakers only

1. Describe how your situation awareness is affected by hearing a non-native English speakers switch between languages

BACK NEXT

Bilinguals only

1. Which languages can you speak? Please order them according to your proficiency, from the L1 – first language, your native language:
   L2 means second language, L3 means third language, etc.

2. Your current proficiency in English language according to ICAO:
   - 4
   - 5
   - 6

3. Please describe how much training you have had in English language?
   How old were you when you started to learn English language:

4. How long have you been trained (years)?

5. Where did most of your training occur?
   - Home
   - School
   - Work
   - Study abroad programme
   - Other:

   How long was your study abroad programme, if any?

6. Have you lived in an English speaking country for more than 3 years continuously?
   - Yes
   - No

   If yes, how long?

7. What is the average ratio of the use of your L1 (native language) and English language (in %) per week?
   L1: English (Together 100%)

8. Please indicate the percentage of time approximately that you use English language each week; for each category separately:
   Reading; Speaking; Writing; Listening

9. While working, do you talk with your colleagues (who speak the same first language as you do) in your native language when you are not communicating by radio?
    - Yes
    - No

10. Have you ever experienced a situation when you struggled to talk in English language?
    - Almost every flight/shift
    - Once a week
    - Once a month
    - Once per three months
    - Once a year
    - Never

11. How would you describe situations in which you struggle to express what you want to say in English language?

12. Please describe how your situation awareness is affected by switching between languages:

13. Have you ever experienced a situation when you probably did not precisely say what you intended to say ("I thought that I said") because of the language alternation?
    - Almost every flight/shift
    - Once a week
    - Once a month
    - Once per three months
    - Once a year
    - Never

BACK NEXT

Common questions for both native-English speakers and non-native English speakers

1. How often have you experienced the use of local languages at international airports in the last year?
    - Very frequently – I am ATC and I use both local and English language every shift
    - Very frequently – I am pilot and I have experienced it every time I fly to international airports (maybe except only few)
    - Frequently – not every time, or every shift, but still the use of the local language is higher than of English language. How would you express this dominance in percentage?
    - Occasionally – the use of the English language is higher than of the local language. How would you express this dominance in percentage?
    - Rarely – it can happen rarely, sometimes more often, sometimes less often, it depends
    - Very rarely – it can happen, but it is very rare

    - Never - please circle whether you are controller or pilot (see the space in the question below)

2. Please fill the additional answer for previous question for options Frequently, or Occasionally, or Never.

3. What kind of problems poses switching between languages for you?

4. What was the consequence of the worst language difficulty you have experienced?
    - Emergency, incident or near miss
    - Wrong task performance – corrected and resolved
    - Prolonged transmissions
    - No consequence

5. Have you ever experienced a situation when you misheard the beginning of the message (addressee of the message) because of a language switch?
    - Almost every flight/shift
    - Once a week
    - Once a month
    - Once per three months
    - Once a year
    - Never

6. In general, how much attention is required for you to understand what a non-native English speaking pilot/controller is saying in English as compared to native-English speaking controller/pilot?

7. From your own experience, please make any other relevant comments you feel may assist this survey:

BACK SUBMIT

4.2.3. Procedure

Data collection was conducted in two runs. First, invitations to complete an online survey were distributed via six aviation forums (Airliners_net, Live ATC net, Pprune, AvCanada, Reddit and Jet photos), between August and December 2015. Additionally, pilots from two airlines (Etihad Airways and Qatar Airways) were asked to participate. The second data collection was conducted between September 2016 and April 2017. The invitations to distribute the survey were sent to three organizations: the International Federation of Air Traffic Controllers' Associations, EUROCONTROL, and the European Aviation Safety Agency. All data were recorded anonymously, and participants were not asked to provide any details that could lead to their identification (e.g., age or gender). The data were analysed quantitatively and qualitatively to allow for the identification of issues related to bilingualism. This project was evaluated by peer review and judged to be low risk. A copy of the institutional low-risk notification can be found in Appendix A.

4.3. Results

4.3.1. General findings

To answer the research questions, the survey data were analysed using quantitative and qualitative analysis. Quantitative analysis was undertaken on a small subset of data from the survey and to support qualitative analysis, which was undertaken on the answers provided on the open questions following thematic analysis rules (Guest, MacQueen, & Namey, 2012).

Overall, 214 pilots and ATCOs (pilots, *n* = 181; 84.58%; ATCOs, *n* = 33; 15.42%) completed the survey. The mean number of logged flight hours of pilots was 9844.83 hours (*SD* = 5938.95), and the mean work experience duration for ATCOs was 12.93 years (*SD* = 11.976). The sample of pilots and ATCOs was further described by the types of flights and ATCO roles based on EUROCONTROL's (2006) study. Of the sample of pilots, 38.12% (*n* = 69) operated on intercontinental routes crossing continental border(s), 35.91% (*n* = 65) operated on international long haul routes crossing two or more international borders, 16.02% (*n* = 29) flew domestic flights, and the remaining 9.95% (*n* = 18) operated on international short haul flights. Most of the ATCOs worked in one of three roles: radar (*n* = 12; 36.37%), tower (*n* = 7; 21.21%), and approach (*n* = 4; 12.12%), with only two on ground (6.06%). Of the remaining eight participants (24.24%), each worked on a different position controlling *en* route, non-radar area control, arrivals, departures, or some combination thereof.

Of the 144 participants who answered questions on the context of English language acquisition, 98 participants (68.05%) acquired their English language skills at school, and 17 (11.80%) participants attended a study abroad programme. Other participants obtained their English language skills either at work (*n* = 13; 9.03%), at home (*n* = 8; 5.56%), or reported another means of English language acquisition (*n* = 8; 5.56%), such as immigration to an English-speaking country or travelling. Of the 135 participants who provided information on how long they had been learning English, the mean duration was almost 12 years (*M* = 11.56, *SD* = 9.67). Of the 144 participants, 53 (36.81%) reported having lived in an English-speaking country for more than three years continuously (*M* = 13.7 years, *SD* = 13.86).

To explore the representativeness of the sample in terms of the variety of native languages of the participants, descriptive data of the language families were summarised in Table 2. Of the largest language family (the Indo-European and Uralic family), 6 sub-families were identified. Because they were large themselves, they were analysed separately. The Latin-Romance sub-family was represented by Spanish (*n* = 18; 35.29%), Italian (*n* = 13; 25.49%), French (*n* = 11, 21.57%), Portuguese (*n* = 7; 13.73%), and Romanian (*n* = 2; 3.92%) languages. From the Germanic sub-family, the largest number of participants had German as a native language (*n* = 10; 43.48%), followed by Dutch (*n* = 8; 34.78%), and Swedish (*n* = 5; 21.74%). The most frequent native language of the Indic sub-family was Hindi (*n* = 15; 71.43%). Three participants (14.29%) had Urdu as their native language, two (9.52%) had Sinhalese, and one (4.76%) had Nepali. The Slavic sub-family was comprised of one Polish-, one Slovak- and one Bulgarian-native speaking participant, and two Russian-speaking participants. A majority of the Afro-Asiatic language family spoke Arabic as a native language (*n* = 14; 93.33%). Two languages belonged to the Dravidian family—Tamil and Malayalam—and were each represented by two participants.

Table 2

Demographic: First Language of Participants by Language Families, and Percentage of Participants Speaking more than Two Languages across Corresponding Language Families

Language families/sub-families		*n*	%	Multilingualism, *n* (%)
Indo-European and Uralic Family	Italic (Latin) Romance	51	37.23	26 (51%)
	Germanic	23	16.79	18 (78%)
	Indic	21	15.33	7 (33%)
	Hellenic (Greek)	8	5.84	6 (75%)
	Slavic	5	3.65	5 (100%)
	Uralic (Hungarian)	1	0.73	1 (100%)
Afro-Asiatic (Arabic, Hebrew)		15	10.95	4 (27%)
Korean		4	2.92	0
Dravidian		4	2.92	3 (75%)
Austronesian (Malay)		2	1.46	1 (50%)
Sino-Tibetan/Sino-Thai		2	1.46	1 (50%)
Altaic (Turkish)		1	0.73	0

Note. *N* = 137

More than half (*n* = 72; 52.55%) of those who stated the languages they could speak were multilingual, speaking two or more languages besides their native language.

4.3.2. Bilingual Air Traffic Environment

Findings showed that English was spoken predominantly on over two-thirds (*n* = 142; 66.36%) of international flights experienced by survey participants. Bilingual air traffic was also experienced frequently or very frequently in 28.97% (*n* = 62) of flights and shifts. Of the total number of ATCOs (*N* = 33), three had never experienced bilingual air traffic operation because of the country they worked in, but eight (24.24%) reported using both English and their local language on every shift.

Both NES and non-native English speakers with ESL (*N* = 128) reported that they had experienced a language alternation situation where they misheard the beginning of a message; 56 (43.75%) once per year, 33 (25.78%) once per three months, 25 (19.53%) once a month, 7 (5.47%) once a week, and 7 (5.47%) on almost every flight or shift. Countries where participants had experienced bilingual air traffic with or without communication difficulties were reported by 63 participants (see Table 3). The participants listed nine different non-native English speaking countries, with China and France together accounting for more than half (51%) the experiences.

Table 3

Survey Participants' Experiences of Countries Using Bilingual Air Traffic Communication (ATC)

	n	%
China	19	30.16
France	13	20.63
Russia	9	14.29
Spain	7	11.11
South America	5	7.94
Italy	3	4.76
Canada (eastern provinces)	3	4.76
Thailand	3	4.76
Latin America	1	1.59
Total	63	

Of the 171 responses from both NES and ESL on whether listening to non-native or native English speakers requires more attention, 106 (61.99%) participants responded that more attention was required for non-native speakers, 42 (24.56%) participants reported that there was no difference, and 23 (13.45%) participants reported that it was harder to understand native English speakers because they can, for example, speak fast and use slang. Of those participants who also provided some details or explanation (*N* = 82; 47.95%), the majority (*n* = 61; 74.39%) reported that it depended on the accent of both native and non-native English speakers. For example, Participant 53, a NES pilot, reported: "*Of course you pay more attention to a non-native English speaker. It is the heavy accent that makes the real difference. For example, the Scottish controllers are much more difficult to understand* *than the Asian controllers*." The second most frequent reason provided was English language proficiency of the non-native English speakers (*n* = 8; 9.76%).

In relation to bilingualism, it was also reported that while on duty, 67.6% (*n* = 96) of ESL participants, both pilots and ATCOs, conversed with colleagues in their native language when they were not communicating by radio. The mean approximate percentage of time that ESL participants used English each week in comparison to their native language (and occasionally other languages as well) was 49.74% (*SD* = 27.57).

4.3.3. Perceived Effects on SA

Opposite effects of language alternation on SA were found for native and non-native English speakers, $\chi^2(1) = 52.167$, $p < .001$ (see Table 4).

Table 4

Perceived Effects of Language Alternation on Situation Awareness of Native- and Non-Native English-Speaking Participants

Non-native English participants			Native English participants		
Category	*n*	%	Category	*n*	%
Not affected	75	68.18	Not affected	5	7.94
Decreased	32	29.09	Decreased	48	76.19
Improved	3	2.73	Not experienced	10	15.87
Total	110		Total	63	

The findings suggest that while language alternation was mostly perceived to cause no effects (68%) on the SA of non-native English speakers ("*Not affected at all. I am so used to it that I hardly notice it*", Participant 103, pilot), the SA of native-English speaking participants was perceived to be adversely affected (76%). For example, Participant 131, a pilot, explained being affected "*dramatically; you become totally reliant on TCAS to understand what is happening around you.*" Participant 12, an ATCO, commented that "*it makes you pay more attention to read-backs and then to be more aware of what they are doing to ensure they follow instructions*," and Participant 48, a pilot, further explained "*the classic example is taxiing across an active runway in low visibility in China… other aircraft* *are given clearance to line up/take off in Chinese. You have no idea what they have been cleared to do.*" Besides impaired SA, three (5%) participants of the NES group, who recognised some adverse effects, explicitly reported the out-of-the-loop phenomenon: Participant 149, a pilot, reported being "*affected greatly. As we are not in the loop, what is being said to the other aircraft around us leaves a big gap in the situation awareness.*"

The adverse consequences of language alternation were mainly observed in terminal areas with high volumes of air traffic communication. Participant 67, a NES pilot, commented that "*this "loss" of situation awareness is important at any time, but becomes more crucial whilst in terminal airspace, as well as on an airport surface movement area, where the traffic can, by nature, be of a higher volume and density, leaving a much smaller margin for error*." Participant 110, ESL pilot commented that a consequence was "*lack of situation awareness for other traffic. This is not necessarily very pronounced en route. But it poses a threat in TMAs*[2] *and CTRs*[3] *or other high-density airspace.*"

Of the non-native English-speaking participants who did not perceive any effect of language alternation on their SA, six (5%) reported that they used only English for communications either because of company Standard Operating Procedures (SOPs) or other (unspecified) reasons. For example, Participant 188, a pilot, stated "*situation awareness is not compromised in any case. Our SOPs and our crews are English only,*" and Participant 104, also a pilot, commented similarly; "*it is not affected since the main operational language in the cockpit is English. It will be used when required even with a person of the same native language.*" Participant 195, an ATCO, commented *"we experience this rarely, speaking other languages in the control room is discouraged*," and Participant 132, a pilot, expressed it with the words: "*I have always used English in my work and for me it is easier than my own language.*"

Table 4 shows that 29% of non-native English speakers also experienced adverse effects of language alternation on SA. Participant 147, a pilot, noted that "*SA is impacted whenever the workload increases. Switching between languages might increase some workload that*

[2] The term TMA means Terminal Control Area, a control area normally established in the vicinity of one or more major aerodromes (ICAO, 2001b)

[3] CTR means Control Zone, a controlled airspace extending upwards from the surface of the earth to a specified upper limit (ICAO, 2001b)

I could use for something 'more important'." When alternating between languages, bilinguals may sometimes insert words in their first language while speaking primarily in English. Participant 100, a pilot, gave the example: "*I sometimes use the English grammar/vocabulary in French and vice versa*." Additionally, ATCOs may issue clearances in a local language to a crew speaking in English. Participant 27, a pilot, commented "*I have also been spoken to by a local controller in the local language—he promptly corrected himself and retransmitted in English*," and Participant 67, also a pilot, responded "*quite often in certain parts of the world, controllers will become confused after speaking their native language to a number of other aircraft and will issue instructions/clearances to English speaking crews in that native language*."

4.3.4. Experienced Consequences

A quarter (*n* = 34; 24.11%) of participants with ESL described having experienced situations in which they had struggled to speak English at least once a month (see Table 5).

Table 5

Reported Frequency of Experienced Struggles to Speak English by Non-Native English-Speaking Participants

Category	*n*	%
Almost every flight/shift	6	4.26
Once a week	8	5.68
Once a month	20	14.18
Once per three months	20	14.18
Once a year	27	19.15
Never	60	42.55
Total	141	

Of the 72 participants who described having had trouble speaking English, 26 reported that they had struggled to converse in English in two major situations. These circumstances included either stressful non-standard situations that contained some kind of urgency while they performed their duties (*n* = 12; 46.15%) (e.g., "*When I am in an emergency situation, I express myself automatically in Spanish even though I should express myself in English* *due to the SOPs*," Participant 187, pilot), or situations of casual conversation about unfamiliar topics, not directly related to their duties or aviation in general (*n* = 14; 53.85%). Furthermore, of the 72 participants, 20 (27.78%) reported that their difficulty speaking, regardless of the context, was manifested in struggles to actively use vocabulary, which was expressed with statements such as "*I cannot find the right word, or I forgot a word*" (Participant 134, pilot). Ten (13.89%) of the participants who struggled to comprehend a conversation considered the different accents of both native and non-native English speakers to be the main cause of their difficulty.

Eight participants (11.11%) reported different strategies they have used to overcome their experienced struggles to speak English. These included "*Use other words or examples*," (Participant 75, pilot) and "*I have to stop talking and think in my native language and try to translate it*" (Participant 55, pilot). Nine participants (12.50%) also reported different emotions experienced when struggling in speaking or comprehending English (e.g., embarrassment, frustration, stress or feeling handicapped). Consequently, flight task performance can be affected (see Table 6).

Table 6

The Consequences of Difficulties Related to Bilingual Air Traffic

Category	*n*	%
Emergency, incident, or near miss	11	5.47
Wrong task performance—corrected and resolved	43	21.39
Repeated transmission	78	38.81
No consequence	69	34.33
Total	201	

Exploration of the effect of language alternation in which participants did not precisely say what they intended to say ("I thought that I said") revealed that of the 142 participants who responded to this question, approximately 60% had either never experienced this effect (*n* = 41; 28.87%) or only experienced it once a year (*n* = 46; 32.39%). Only a small number of participants (*n* = 4; 2.83%) reported experiencing such a situation on almost every flight/shift, but a trend swiftly increased for weekly periods (*n* = 12, 8.45%) followed by small increase in monthly periods (*n* = 15, 10.56%), and the occurrence doubled between weekly and quarterly periods (*n* = 24, 16.90%).

The following two examples illustrate the potential for confusion over clearances, as reported by native-English speakers:

"*Worst I ever dealt with was several flights of Singapore air force pilots who repeatedly took control instructions issued to different flights. I would issue one flight a turn and three other flights would turn as well*" (Participant 31, ATCO).

"*Not sure when a clearance has been read back... so you are not sure when their conversation has finished. I don't want to interrupt a transmission*" (Participant 119, pilot).

A hazard potentially inherent in a bilingual air traffic environment is that pilots who do not understand the local language may not pay attention to that conversation. Three examples are provided for illustration:

"*Loss of SA, particularly as it relates to the location of other aircraft. Turning down the radio as transmissions are more of a distraction*" (Participant 27, pilot).

"*As pilots we almost all have a tendency to 'tune out' non-essential transmissions... When switching between listening to clear and expected communications, to transmissions that require additional thought process, I am sure that less emphasis is placed on the more difficult*" (Participant 76, pilot).

"*For example, when in Moscow and the controller is speaking in Russian, I tend to not listen as closely, knowing that I cannot gain any information from the controller at this moment. As he finishes speaking to this particular aircraft I again have to remember to shift my attention back to the controller to see which language he will speak and to which aircraft he will speak... This example of switching languages seems to happen at the airports that also have more complicated arrivals. They make a difficult situation more difficult by switching languages and add a greater element of risk where there would not have to be one*" (Participant 86, pilot).

Another area of bilingualism in aviation—communication between pilots and tug drivers, or maintenance (ground personnel)—was identified: "*I worked in Hong Kong for 27 years and all tows on the airport used Cantonese and approvals and instructions were given in English to an ATCO Assistant, who then passed the instructions to the tug driver. This meant that pilots had no idea what was going on around them with regard to tows*" (Participant 204, ATCO).

4.4. Discussion

To explore how the situation related to bilingual air traffic communications has changed since English LPRs were introduced, and what the current challenges are, an online survey was distributed to both NES and non-native English-speaking (ESL) pilots and ATCOs. There was some evidence that pilot and ATCO proficiency in English had improved. Only 13% of bilinguals attained the minimum level 4 (Operational) requirement, with the remainder achieving higher levels of proficiency. Despite this, bilingual air traffic was experienced frequently or very frequently in 31% of flights and shifts. Although English was spoken predominantly on a radio, 80% of NES participants, who did not understand a local language, reported decreased SA that they attributed to the use of two languages on the same radio frequency. Somewhat surprisingly, almost 30% of bilinguals also reported that alternating between the languages could increase their workload. It might be expected that language alternation would cause no effects on their SA, given that they can understand both languages that are used for broadcasting. This may suggest that the SA of bilinguals is not necessarily limited to the simple comprehension of two languages but might be related to their alternation as well. In other words, language alternation might have an adverse effect on the SA of bilinguals, even when pilots and ATCOs can speak both languages used for air traffic communications.

Difficulty speaking English, regardless of any particular situation, was manifested in struggles to actively use vocabulary, and was expressed with statements such as "*I cannot find the right word, or I forgot a word*" (Participant 134, ESL pilot). These findings are in accordance with those of previous studies (Buchanan, Laures-Gore, & Duff, 2014; Grosjean, 1982; Murray, Baber, & South, 1996; Saslow et al., 2014; Tajima, 2004; Orasanu, 1997), which found that even with satisfactory English proficiency, unusual and high workload situations may reduce the ability of non-native English speakers to communicate in and comprehend English. For example, pilots involved in the Tenerife accident had a high command of English, yet, the KLM captain code-switched (Tajima, 2004); that is, he used the grammar of his native Dutch language while speaking English. This is not really surprising, given two factors; first, both languages are still active in the brain and compete with each other for selection (Meuter & Allport, 1999), and second, it has been found that in situations of perceived stress, the language areas in the brain are impaired (Saslow, et al., 2014).

There are two main areas in the brain related to language processing: Broca's and Wernicke's area. The former is responsible for language production, and the latter for understanding language (Servan-Schreiber, 2004; Ye & Zhou, 2009). During stress, the body releases hormones, which potentially block the function of Broca's area (Servan-Schreiber, 2004). This potentially explains why people who have experienced great fear or are stressed, can be speechless or may stutter, and also why bilinguals may understand what was said, but struggle to speak. This description of the process of language perception and production is certainly only a sketch of what neurolinguistics have to offer about this very complex process. Nevertheless, it entails the basic processes that can help to understand why sometimes even very proficient ESL aviation personnel struggle with communication, without necessarily ascribed them a lack of English proficiency.

In aviation, non-routine situations can present flight-related stressors. This type of situation places larger demands on communication skills, requiring pilots and ATCOs to use plain language where standard phraseology may not be sufficient to describe a situation (ICAO, 2001a). Based on the effect of stress on language processing in the brain, as mentioned above, it can be hypothesised that particular reactions can make such situation worse. For example, when a crew with ESL experience a non-standard situation, they may struggle to describe the situation to an ATCO. By remaining calm, the ATCO can have a calming effect on the crew. Calm communication can facilitate the desired goal of mutual understanding. In contrast, an impatient or verbally aggressive reaction from the ATCO could increase the stress of the crew, and thus, adversely affect their ability to describe the situation.

Another difficulty reported in relation to bilingual air traffic was code switching— where the participant inserted a word from their native language into an English sentence or issued a clearance in a local language to an English-speaking crew. Code switching has been described in a number of studies (e.g., Gollan, Sandoval, & Salmon, 2011; Grosjean, 1982; Hermans, et al., 1998). For example, it was found that deviations from standard phraseology in the presence of Italian language in radio transmissions increased during nightshifts, when the workload was low, while the most correct exchanges occurred in a high workload shift (Corradini & Cacciari, 2002). Additionally, Borins (1983) also argued that ATCOs are most prone to error, not when traffic is at its heaviest peak, but when it begins to decrease and ATCOs start to relax. These results suggest that the circumstance of lower workload decreases radio discipline and, therefore, can provide a room for errors. Tiewtrakul and Fletcher (2010) found that communication errors occur significantly more often when speakers are both non-native English speakers.

The data from this study provided some evidence that SA was mainly affected in the terminal area, where there is larger volume of radio communication. Given that bilingualism appeared to affect SA most often during high volumes of radio communications, when radio discipline was found to be at its best (Corradini & Cacciari, 2002), it may be the alternation itself that affects SA, rather than the command of English. In other words, frequent switching between languages is more prone to decreased SA in bilinguals even though they are highly proficient in English. On the other hand, language alternation occurrences were found to be more frequent during low workload shifts (Corradini & Cacciari, 2002), and thus, should not pose a large threat to bilinguals' SA.

Some adverse effects on emotional experience were also recognized. Habrat (2013) and Piasecka (2013) discussed the relationship between language use and affects, and self-esteem. It was found that the fear of negative evaluation made participants feel anxious and often resulted in avoidance of speaking and thus missing the opportunities to practice the verbal presentation. Having to perform in a language in which they could not express themselves properly potentially caused non-native speakers' communication apprehension. The communication breakdowns *per se*—without being exposed to additional stressful situations of emergency in the aviation context—evoked anxiety and confusion. Lacking words to express ideas led to various emotions from anger to depression.

Training ESL pilots in English-speaking countries can improve the language acquisition process (Xinhua News Agency, 2007). This trend seems to be relatively new given the data of this study, which indicated that most of the participants (*n* = 98; 68%) acquired their language skills at school. However, there is not sufficient empirical evidence to support this assumption, because some relevant demographic questions were not included in the survey, such as the age of the participants when they started learning English, and the year they obtained their professional licences. Despite this, living in an English-speaking country may be help to solve language acquisition issues. However, one participant indicated possible problems with ESL student pilots training in an English-speaking country, which was consistent with Estival and Molesworth's study (2009): "*I work at an aerodrome with many different international flying schools. I have had many occasions where a clearance wasn't understood but was read back and the student did what they usually do rather than what was cleared. Or they may not reply at all when given a clearance they don't understand*." (Participant 16, ATCO).

Importantly, while on duty, 67% of ESL participants communicated in their native language when they were not communicating by radio. This may suggest that regardless of the language spoken in air–ground communications, even if it were only English, language alternation will most likely still be present. It is unclear whether performance in a bilingual environment is less adversely affected when English is used more frequently, than when it is used infrequently for radio broadcasts; that is, the use of a local language dominates. There was evidence that some airlines or ATC providers encourage the use of English for all communications, even between colleagues in a cockpit or control room (see section 4.3.3). This may suggest that ESL aviation personnel are somewhat open to a monolingual English air traffic system. This would be significant for NES participants, whose SA decreased (*n* = 48; 76%) when the two languages were spoken in the same airspace. Besides the out-of-the-loop phenomenon, they also reported struggling to make a judgement about when a radio communication ended. Not knowing when a communication has ended was identified as a bilingualism factor contributing to blocked transmissions (EUROCONTROL, 2006).

The most frequently reported countries where participants have experienced bilingual air traffic, with or without struggles in communication, were China (*n* = 19; 30.16%), France (*n* = 13; 20.63%), Russia (*n* = 9; 14.29%), Spain (*n* = 7; 11.11%), and South America (*n* =

5; 7.94%). It must be noted that the frequency of reports could be relative to the frequency with which international flights were conducted to/from each country. Unfortunately, this information is unknown from the obtained data. Moreover, there were no questions about which countries the communication problems were experienced in.

There were three potential limitations to the study reported here. First, after the survey had been distributed, it was realised that some grammatical mistakes were made in the survey. The original wording of the survey that participants read can be seen in Section 4.2.2.1. (mistakes were retained to demonstrate this limitation). That said, it was believed to be unlikely these actually affected participants' understanding of the survey. Second, some key terms, such as language switching and situation awareness, were not defined at the beginning of the survey form. In hindsight, it was determined that the absence of definitions of the terms might have caused ambiguity and adversely affect the participants' responses. Third, Questions 7 and 8 of the survey for bilinguals were not included in the analysis, because it was realised, in hindsight, that they were ambiguous. The responses on these questions revealed that the questions were not clearly formulated and might be understood differently between participants. Question 7 was: *What is the average ratio of the use of your L1 (native language) and English language (in %) per week?* Question 8 was: *Please indicate the percentage of time approximately that you use English language each week; for each category separately: Reading; Speaking; Writing; Listening.* Additionally, responses on these questions were subjective estimations, rather than objective. This method of a self-rated frequency of using English, even though not precise, was used in a study by Zheng, Roelofs, and Lemhofer (2018). Moreover, Questions 7 and 8 asked for similar information.

Question 7 did not specify whether the environment was at work only or in general, which may have caused discrepancies in the answers provided. Question 8 was answered in two ways. Some participants answered the question in terms of the percentage of time they used English (rather than another language) for each task. For example, 80% of their reading was in English (with 20% in another language), along with 20% of their speech, 5% of their writing and 40% of their listening, giving a total of more than 100%. Other participants assumed the division of their time into the four categories should yield a total of 100% (e.g., Reading: 20%, Speaking: 40%, Writing: 10%, Listening: 30% = 100%). In the former approach to the question, participants compared each of the language related activities—

reading, speaking, writing, and listening—between the languages they used (English and their native language or languages), whereas in the later approach they compared the proportion of all four *activities* when using English.

These limitations have led to a lesson for the future design of survey questions. The questions must be very clear, because there is no opportunity to clarify their meaning with participants. Prior to the data collection, a pilot study should be utilized to see how the questions are answered, and to confirm that they are understood in the way intended by the research objectives.

To conclude, a question was raised as to whether bilingual air traffic is still somehow beneficial for those who can alternate between languages when the English language proficiency of bilinguals has improved. Based on the findings of this study, even if radio communication were limited to only one language, the problem with language alternation is unlikely to disappear completely. This supports the need to look at language alternation from a cognitive perspective rather than from the perspective of any specific bilingual environment situation.

CHAPTER FIVE

Study 2: Call Sign Recognition

5.1. Introduction

To explore the first level of SA, call sign recognition will be explored, as accurate recognition of a call sign—used to identify an aircraft—is a necessary precondition for every successful communication over the radio channel and to initiate an appropriate reaction. To contact and communicate with a particular aircraft, an ATCO issues that aircraft call sign followed by a message. The pilot of the intended aircraft [ideally] hears the message, recognizes that the call sign belongs to the aircraft being flown, comprehends the message, decides how to proceed, and then verbally responds. In the literature review, and in Study 1, it was revealed that bilingual air traffic can sometimes cause pilots to misperceive their own call sign. This study, therefore, aims to empirically compare performance speed and accuracy when recognizing a call sign between native (L1), and second language (L2) conditions, and a bilingual condition (Mix), in which native and second languages randomly alternate. By doing so, the findings may provide empirical evidence of the language condition that can facilitate faster and more accurate recognition responses.

The first guiding research question for Study 2 was:

Question 2a: Do speed and accuracy of call sign recognition differ between monolingual and bilingual conditions?

5.1.1. Call Sign Aviation Regulation and Research

The regulations regarding call sign designators and associated telephonies may be found in ICAO Annex 10 (2001a) and ICAO Doc 8585 (2016). Call signs can be numeric, consisting of numbers only (e.g., 123) or alphanumeric, consisting of number(s) followed by one or more letters (e.g., 123H) (EUROCONTROL, 2006). According to the United Kingdom Civil Aviation Authority’s (CAA, 2000) Aircraft Call Sign Confusion Evaluation Safety Study

(ACCESS), of the 482 reports of call sign confusion occurrences received during 1997, the majority (84%) involved only the numeric type of call signs. The use of similar call signs on the same radio frequency may increase confusion and thus give rise to flight safety incidents. For example, transmitting an instruction to call sign AA588 directly after Comair588 on the same radio frequency caused confusion on some radio calls (Denti, 2011), and adversely affected pilots' SA.

Similar call signs explained 33% of the communication problem occurrences in a one-year period from March 2004 to April 2005 (EUROCONTROL, 2006). The ACCESS study (CAA, 2000) revealed, for example, that 27% of misunderstandings involved call signs with the same numbers in exactly the same order (eg., 371 and 371), 13% of misunderstandings involved call signs with the same characters in a different order (anagrams) (eg., 371 and 317), 41% involved call signs ending with the same two digits (e.g., 425 and 325), and 23% involved call signs ending with the same single digit (e.g., 126 and 576). Interestingly, digits "five" and "nine" were confused more often with each other than with other digits (Green & Swets, 1988) (for a more detailed discussion about number processing see section 3.7).

The study into confusion of numeric and alphanumeric types of aircraft call signs conducted by Cox and Vinagre (2004) differentiated between *perceptual* and *cognitive confusions*, with the former representing perceptual processing related with poor articulation, noise, or hearing on the side of a receiver of a message, and the latter concerned with cognitive processing; that is, limitations of short term-memory or expectation bias. Two types of stimuli were used in their study: a set of triple digit numbers was used to represent numeric call signs and a set of a single letter from aviation alphabet followed by a single digit were used as alphanumeric call signs. The findings showed that the numbers were the most confusing part of the call signs (about 8–10 times more likely than letters) (Cox & Vinagre, 2004). Digits that were the most misperceived were "one" with "nine", "four" with "five", and "two" with "eight." The least misunderstood digits were "zero", "six", and "three." That said, the analysis was done by phoneme-pairs comparison, which seldom occurs as a stimuli in real-life aviation. The second goal of Cox and Vinagre's (2004) study was to examine whether recognition failure was influenced by the position of a digit in the phrase. They found that the middle digit of the three-digit number was significantly more likely to be misrecognised than the first and third digits.

Overall, the review presented in this section supported the use of analysis of numeric call sign similarity in different language conditions. Therefore, the second guiding research question for Study 2 was:

Question 2b: Does call sign similarity affect call sign recognition in monolingual and bilingual conditions?

5.2. Method

5.2.1. Overview

A computer-based experiment was developed, in which participants were required to identify target call signs among distractors. Participants were Chinese native speakers, with ESL. Chinese participants were used because China is one of the countries that provides bilingual air traffic services (Dennis, 2015b).

5.2.2. Participants

Participants were 42 non-aviation Chinese–English bilingual students enrolled at Massey University, New Zealand. Eight participants were excluded from the data analysis because their native language was considered to be English, with Chinese considered as a second language. Data from 34 students (19 males and 15 females) with Chinese as their native language (both Mandarin and Cantonese dialects) were analysed. The mean age of the participants was 23.94 years (*SD* = 4.87). The mean duration of their stay in New Zealand was 1.97 years (*SD* = 1.93; *Range* = 1 month–10 years). Participants were assigned to one of two English language proficiency groups (low vs. high) according to the IELTS test scores they reported having achieved immediately prior to commencing their studies. No participant reported having any known hearing impairment.

5.2.3. Design

5.2.3.1. Overview

The primary aim of this study was to test the effect of language switching on call sign recognition. To do this, a 3 × 3 × 4 within-subjects experimental design was used with the three within-subjects factors: Language condition, Inter-stimuli interval, and Similarity (the degree to which the three-digit target and the distracting stimuli were similar). A secondary aim was to test the effect of the between-subjects factor of language proficiency (high vs. low ability) on the performance in English language and the Mix condition, for which a one-way between-subjects analysis of variance (ANOVA) was used.

A factorial design is a common method used in aviation psychology language-related research (e.g., Barshi & Farris, 2013; Estival, Farris, & Molesworth, 2016). The main benefit of using this methodology is that it allows flexible control over variables; that is, the analysis and provision of accurate estimates of the effects of independent variables with several levels, and thus comparison of the means of all measurements for examined conditions (Seltman, 2015).

The main shortcomings of this type of methodology have been discussed in Section 3.8, and mostly relate to conducting this type of research in real-life conditions. This study investigated the effect of different language conditions on the performance of bilinguals who can speak the languages used in an experimental task. For a number of reasons, conducting this type of research in the real-world environment with aviation personnel was considered to be unviable. First, because behavioural observation of the cognitive processing of language would not be possible (Zwitserlood, 1998). Second, the availability of Chinese – English bilingual pilots in New Zealand was found to be very limited. Third, the possibility of travelling and staying overseas (e.g., to China) in order to recruit bilingual participants was precluded due to a lack of available funds. As such, it was concluded that the chosen methodology was an acceptable, taking into account the three reasons mentioned previously.

5.2.3.2. Development of the Experiment

To explore the effect of language condition on performance speed and accuracy, three conditions were developed: a pure Chinese language condition (L1); a pure English language (L2); and a language switching condition (Mix) composed of English and Chinese stimuli. These were designed to be analogous to a monolingual air traffic environment rather than a bilingual one, even though the L1 condition would not be experienced as frequently as the Mix or the L2 condition (at least not at international airports). Consistent with the concept of language inhibition of momentary irrelevant language (e.g., Meuter & Allport, 1999), the two monolingual conditions (English or Chinese) did not follow each other, to avoid any performance difference caused by persisting inhibition of irrelevant language in the preceding condition. Because the language which is not currently used seems to be inhibited (Meuter & Allport, 1999), it was expected that the activated language of the preceding condition would interfere with the following condition in different, previously inhibited, language.

The experiment was created using the open-source application PsychoPy 1.82.01 (Peirce, 2007), in which the order of presentation of the stimuli in each language condition was set randomly to minimize the potential for any learning effects. The experiment was designed using the guidelines provided by the Massey University Ethics Committee and was deemed to be low risk by a peer review. A copy of the institutional low-risk notification can be found in Appendix B.

5.2.3.3. Measures

The dependent variables were the RT and the type and number of errors. Two types of error were analysed, misses (e.g., a target number was presented but participants missed it by responding to it as they would to a distracting stimulus) and false alarms (e.g., participants responded to a distracting stimulus as if it were a target number). These dependent variables are an important element of this thesis as they allow predictions about the experimental language conditions that would elicit faster and more accurate responses.

Three within-subjects independent variables were investigated, Language condition, ISI and Similarity of a call sign. The Language condition factor had three levels: the target and distracting stimuli were presented either in Chinese (L1) only, English (L2) only, or in a language switching condition (Mix).

In each of the language conditions, the ISIs of 1, 4 or 9 s were randomly distributed. ISIs of 1 s and 4 s each appeared 17 times per language condition and the ISI of 9 s appeared 16 times in each condition.

The definition of the Similarity factor was based on the idea of identical word-monitoring (Marslen-Wilson & Tyler, 1980), where participants listen to words for the occurrence of a target specified in advance. Prior to the experimental task, participants know which target they must listen for. This can be analogous to pilots knowing their call sign prior to a flight and maintaining their radio for any uses of that call sign by, for example, an ATCO. Because the confusion of the two call signs is based on their similarity, given by a partial match between target and distracting stimuli, construction of this factor can be based on fragments of the pre-known target. For example, in the case of the three-digit target numbers, there are three fragments, the three digits. The recognition of the target was, therefore, understood as a process of piecemeal identification of the digits identical to the pre-known three-digit target. This method of construction of the Similarity factor can provide information as to how much speech information is sufficient for participant to recognise the target call sign and can help to explain the nature of call sign confusion.

The Similarity factor had four levels, and aimed to capture the extent to which an acoustically presented stimulus was similar to the target stimulus. For example, if participants were instructed to identify the target stimulus *531*, a Similarity level of ‘0’ meant that the presented number was completely different; that is, it had no identical fragments (e.g., 125). A Similarity level of ‘1’ meant that the digit in the first position of the presented number was the same as in the target stimulus (e.g., **5**64), but the remaining two digits were different. A Similarity level of ‘2’ meant that the two digits in positions one and two were the same as in the target stimulus (e.g., **53**8) and only the digit in the last position was different from the target number. For statistical analysis, this method of similarity development was also followed for the target number which has to be presented within the stimuli. That is, the target number itself was designated as Similarity level ‘3’, because all three digits and their position in the number were identical to those in the target number (e.g., **531**). However, for the readability of the findings this level was named ‘Target’. No stimuli contained any double digits (e.g., 5**11**).

This method was chosen to facilitate a comparison of the current findings with the outcomes of the ACCESS study (CAA, 2000), which revealed an increased adverse effect caused by increased similarity of two call signs *ending* with the same digits (not starting with the same digits as in the present study). For example, the similarity of level ‘1’, in the ACCESS study, would meant that first two digits were different and only the last single digits in the two call signs were the same (e.g., 14**3** and 52**3**), and level ‘2’ would meant that two call signs had different first digits and the same two digits at the end (e.g., 1**23** and 5**23**). The aim was to explore whether a similar effect would be observed when the similarity increase began from the first digits in the numeric call signs.

Participants’ ability in English language formed a between-subjects independent variable, represented by self-reported IELTS Listening scores obtained before entering a study programme at Massey University. A more detailed description and consideration of the IELTS test can be found in section 5.2.4.2.

5.2.4. Materials

5.2.4.1. Acoustic Stimuli

Participants were assessed on how quickly and accurately they responded to acoustic stimuli designed to represent numeric call signs. This task was designed to represent level 1 SA. Consistent with the ICAO Safety Advisory (2015), numeric call signs were limited to a maximum of three digits. All stimuli were spoken by a computerised female voice (using the OS X Text-to-Speech programme) and recorded over a white noise background (using Audacity 2.1.0). The decision to use a female voice was determined by a limitation of the OS X Text-to-Speech programme, which provided only a female voice option for Chinese language. Fortuitously, previous research has reported no difference between the effects of male or female voices on judgements of perceived urgency of cockpit warnings (Arrabito, 2009); similarly, acoustic and non-acoustic differences between male and female speakers were found to be negligible (Edworthy, Hellier, & Rivers, 2003). A speech to noise ratio (SNR) was chosen to maximize intelligibility of the stimuli and, therefore, neither speech signal nor noise were amplified. White noise and the level of SNR were chosen following the rationale to test a simple model before dealing with additional complexities in the latter studies.

For stimuli pronunciation, two documents about the transmission of numbers were considered. According to ICAO's Annex 10 (2001a), almost all numbers should be transmitted by pronouncing each digit separately (e.g., 238 should be "two three eight"). In contrast, the FAA's Order JO 7110.65W (2015) states that call signs comprising a series of digits are pronounced in group form; that is, as the whole number or pairs of numbers they represent (e.g., 238 as "two thirty-eight"). The idea of this method can be based on a chunking strategy by which individual digits are grouped into fewer units reducing the constraints of short-term memory (Prinzo & Morrow, 2002). A pilot then does not have to remember three units (three digits), but two chunks, which can facilitate their memory. This study was, however, not primarily focused on the memory. Instead, it was decided that each stimulus would be pronounced as a hundred number (e.g., 238 as "two hundred and thirty-eight"), because hundred and thousand numbers were found to have been confused in some aviation incidents (Cushing, 1994).

The target number remained the same within each language condition but differed across conditions. Thus, there were three different targets each presented 16 times in their corresponding language condition. In the Mix condition, the target stimulus was presented in both languages; 8 in Chinese and 8 in English. The 16 target stimuli were randomly distributed within 34 distracting stimuli, giving 50 stimuli in total in each of the three experimental conditions. The probability of a distracting stimulus being presented was 68%, and the probability for a target was 32%. Over the three experimental conditions, each participant was therefore presented with 102 distracting stimuli and 48 target stimuli. The higher proportion of distracting stimuli was designed to simulate real-life aviation communications, where infrequent target stimuli are interspersed with more frequent distracting stimuli not directly relevant to a pilot. Distracting stimuli varied across and within the conditions.

Finally, according to Marslen-Wilson's (1985) finding that words can be processed before they are fully heard, it was possible for participants to respond to a stimulus while it was still playing. The RT (in seconds) was therefore measured from the onset of the auditory stimulus, and the mean RT then calculated and further corrected following the rationale proposed in section 3.12 to obtain the mean pure RT. There were no visually presented stimuli[4], and no other acoustic stimuli (e.g., tone introducing the change between language conditions). The stimuli list is included in Appendix C.

5.2.4.2. English Language Proficiency

English language proficiency was measured using the International English Language Testing System (IELTS). This section describes the rationale for this choice, the test itself, and considers its reliability and validity in the context of the thesis.

One of the eligibility criteria for entry into a study programme at Massey University is measurement of proficiency in the English language. At present, one of the most widely used, recognized and accepted tests for this purpose is the IELTS (British Council, 2018). IELTS is designed to assess the language ability of candidates who want to study or work in English-speaking countries (Exam English Ltd., n.d.).

Prior to enrolment in a degree course, prospective students must obtain a minimum overall score of IELTS 6.0 (Massey University, 2017). For the Pre-Degree and Foundation Certificate the minimum requirement is IELTS 5.0. Some programmes require a higher level of English Language competency than the minimum requirement indicated. Doctoral degree candidates must obtain a minimum English Language competency level of 6.5.

Normally, the IELTS assesses four skills: Listening, Writing, Reading and Speaking (British Council, n.d.a). As the studies of this thesis were focused on language perception rather than production, only proficiency in the IELTS Listening test was assessed. The IELTS Listening test uses a variety of voices and native-speaker accents. It is broken down

[4] In principle it might have been possible to have presented numbers visually using Chinese logographs and Arabic numerals, however, this was not done in the present study as it would have no relevance to aviation, where communication is mostly presented acoustically.

into four sections, each with an increasing level of difficulty. This allows a wide range of listening skills to be assessed, including understanding of main ideas and specific factual information, recognising the opinions, attitudes and purpose of a speaker and following the development of an argument (British Council, n.d.c).

Before the use of the IELTS Listening test, its validity and reliability was considered. Academics and researchers worldwide who use the IELTS report yearly on the distribution of scores achieved in various contexts (i.e., test takers by country or region) (IELTS, n.d.a), and its reliability estimates (IELTS, n.d.b). To consider the reliability of the use of the IELTS in the context of this thesis, the IELTS reports were reviewed. Some studies tested populations of Chinese students (e.g., Badger & Yan, 2012; Lloyd-Jones & Binch, 2012; Wray & Pegg, 2009), and some studies were conducted in New Zealand contexts (e.g., Merrifield, 2012; Read, Wette & Deverall, 2009; Smith & Haslett, 2007, 2008). Some studies were focused particularly on the IELTS Listening test (e.g., Breeze & Miller, 2011; Coleman & Heap, 1998; Winke & Lim, 2014), and, specifically on Chinese students performing IELTS Listening test (Badger & Yan, 2009). Based on these studies, it was assumed that the IELTS Listening test could be used as a valid and reliable tool in the context of the studies of this thesis. The assumption can also be supported by the fact that the study participants were studying at Massey University in New Zealand, which is the target population for which the IELTS test was developed. Therefore, IELTS was recognized as a suitable test despite its potential limitations as is explained below.

One of the limitations of the IELTS that needed to be considered prior the experiment is that performance in IELTS can be affected by preparation for and familiarity with the test. Participants had completed the IELTS prior entering their study programme at Massey University. However, even if participants had sat the IELTS again specifically for this thesis, their previous experience with IELTS would not have affected their performance, because it had occurred approximately 2 years previously. This is sufficiently long period to prevent any ‘practice’ effect on their performance. Therefore, it might be expected that this factor would neither adversely, nor positively, affect their performance.

Because participants had already conducted the IELTS, and because the task in Study 2 did not put large demands on English language skills, the participants were not retested but, instead, reported the scores they had obtained at their entry into the Massey University study programme. The IELTS scores for English proficiency are considered to be valid for two years from the test date, suggesting that the self-reported information would be a reasonably valid measure of their English language ability at the time of participation. A similar method has been used in previous studies (e.g., Koch, Decklerck, & Philipp, 2015; McClain & Huang, 1982; Marsh & Maki, 1976; Zheng, Roelofs, & Lemhofer, 2018).

In summary, the decision to use IELTS allowed comparison of findings for Studies 2–5. Based on the information presented herein, it was estimated that the IELTS Listening test could be considered as valid and reliable tool (see also Aryadoust, 2013) to measure English language proficiency of participants for this thesis and in this environment.

5.2.4.2.1. ICAO Rating Scale and IELTS Alignment Approach

An important element of this thesis was to make assumptions about the English language proficiency of aviation personnel from the findings of the thesis. This required at least an approximate alignment of the scores of the IELTS with the proficiency levels according to the ICAO Rating Scale.

The IELTS tests and ICAO recognized tests are different; they have different structures and serve different purposes with different target populations. Moreover, for this thesis, only the Listening part of the IELTS test was used. The ICAO rating focuses on listening and speaking, and more importantly, uses technical vocabulary for aviation-specific content (Harcourt Assessment, 2006). It scores Pronunciation, Structure, Comprehension, Vocabulary, Fluency and Interaction (ICAO, 2010), not all of which are covered in the IELTS Listening test.

Nevertheless, the approximate relations between IELTS scores and ICAO levels were made possible by comparing two studies that used the Common European Framework of Reference for Languages (CEFR) levels as a common reference of English proficiency levels. One of the studies (Cambridge English, 2016) compared IELTS with the CEFR levels, while a study by Harcourt Assessment (2006) compared CEFR with the ICAO Rating Scale. This made approximate alignment of the IELTS with the ICAO scale possible (see Table 7).

Table 7

ICAO Rating Scale, CEFR, and IELTS Comparison Chart

ICAO Rating Scale	CEFR	IELTS
	B1	4.0–5.0
Operational level 4	B2.1	5.5–6.5
Extended level 5	B2.2	
Expert level 6	C1	7.0–8.0
	C2	8.5–9.0

According to Cambridge English's (2016) report, an IELTS score of 5 is comparable to CEFR B1. IELTS scores between 5.5 and 6.5 can be aligned to CEFR B2 (B2.1 and B2.2) with 5.5. being the borderline between B1 and B2.1. Candidates who are at level C1 of the CEFR scale can be expected to be comparable in ability with candidates who have secured 7.0 or 8.0 in IELTS. However, an IELTS score of 8.0 is borderline between CEFR C1 and C2. Finally, the CEFR C2 level is comparable with an IELTS score of 8.5–9.0. According to Harcourt Assessment (2006), ICAO Operational level 4 can be equivalent to CEFR level B2.1, Extended level 5 is comparable to CEFR level B2.2 and Expert level 6 is roughly comparable to CEFR level C (C1 and C2) for Comprehension and Vocabulary skills, which are measured by IELTS Listening as well.

Together, the minimum required level for aviation personnel, ICAO level 4, is roughly comparable to IELTS level 6.0. Importantly, this is just to allow making orientation estimations about the ICAO language proficiency rating scale from the findings in this thesis as a matter of discussion; it does not mean that the tests are equivalent.

5.2.4.3. PsychoPy

A search was conducted for software that allowed development of a computer-based experiment using acoustic stimuli and measurement of both response speed and accuracy. Several options were investigated including e-Prime™, DMDX, SuperLab, and PEBL. Ultimately, the experiment was created using a free open-source application, PsychoPy 1.82.01 (Peirce, 2007), which, on balance, appeared to best suit the needs and requirements of this study.

The main perceived benefit of PsychoPy was its user-friendly interface consisting of Builder and Coder views. The Builder view is intuitive, easy and flexible to work with, allowing the insertion of stimuli and manipulation of the environment to customize the experiment to individual needs. It allows various types of stimuli and responses. The Coder view requires some programming skills but allows the development of highly professional experiments. PsychoPy uses Python programming language for coding scripts that are designed to be easy to understand, and therefore read and write, even for users with only rudimentary knowledge of programming.

Although programming in PsychoPy requires only rudimentary programming skills, online support, demonstration codes, and tutorials are available online. A worldwide community of users also contribute to PsychoPy's continuous evolvement by sharing scripts and demonstration codes, modifying existing scripts and submitting them back into the package so that the whole community benefits.

5.2.5. Procedure

Participants were recruited either by an e-mail invitation via the Chinese Student Club at Massey University, or personally on the Massey University campus between mid-September and mid-November 2015. Prior to the experiment, participants completed a brief demographic questionnaire that included questions about the IELTS Listening test score they had achieved before entering their study programme, and about any known hearing impairment. The experiment was conducted anonymously—participants were not asked to provide their names.

None of the participants reported any known hearing impairment. Based on their self-reported IELTS English language proficiency test scores, they were assigned to one of two proficiency groups in accordance with the IELTS categorization for the purpose of data analysis (British Council, n.d.b): Modest/Competent users (*n* = 18; IELTS test scores: 5.0–6.5), and Good/Very good users (*n* = 16; IELTS test scores: 7.0–8.5).

To test call sign recognition, participants were provided with a target number and instructed to listen carefully and press "Yes" on a keyboard when they heard the target number and

"No" for any other number. Participants had time to practice the task in a practice block. When they fully comprehended the task, they participated in the experiment. The experiment lasted approximately 20 min. All instructions were provided in English. Participants were provided with refreshments to the value of $5 as gratitude for their participation in the experiment.

A priori power analysis for an omnibus three-way ANOVA was conducted using the software G*Power (Erdfelder, Faul, & Buchner, 1996). Of relevance was specifying the power of the key comparison between three language conditions rather than reporting power analyses across the entire data set for each main effect and interaction. A total sample size of $n = 28$ was recommended for an experimental power of .80, with $\alpha = .05$ and an effect size of $f = .25$ (based on Ison, 2011) in a repeated-measures, within-subjects design. *A priori* power analysis was also used to determine a total sample size in a between-subjects design, for the between-subjects English language proficiency factor, with $\alpha = .05$ and an effect size of $f = .55$ (based on that reported in analyses by Barshi & Farris, 2013). A total sample size of $n = 32$ was recommended for experimental power of .80.

5.3. Results

5.3.1. General Findings

The data were initially screened to identify any outliers, which were defined as z-scores greater than |3.29| as recommended by Tabachnick and Fidell (2007). This screening identified 9 outliers in the language, ISI and similarity data. However, before excluding any outlier, the nature of these observations was considered to decide whether they might be legitimate; the sample represents a population, and it is not unreasonable to expect a certain probability of people with very fast or very slow performance (i.e., the upper or lower extreme scores). To do this, normal Q-Q plots were reviewed, although this can be considered somewhat subjective. There were three major benefits of this approach. Normal Q-Q plots allow us to see whether the assumption is plausible, how the assumption was violated and what data points contributed to the violation (Ford, 2015).

As shown in the example presented in Figure 3, the presence of an outlier in the data of this study would probably not affect the slope of the regression line, which defines the linear relationship between the variables, and can be used to estimate an average rate of change. It would only have an effect on the standard error of the slope estimate (Friedman, n.d.). In other words, although an extreme value of this data set, it does not seem to affect the nature or character of the relationship between the variables, and hence findings. Rather, it increases the data range. Therefore, all outliers that followed this reasoning were retained.

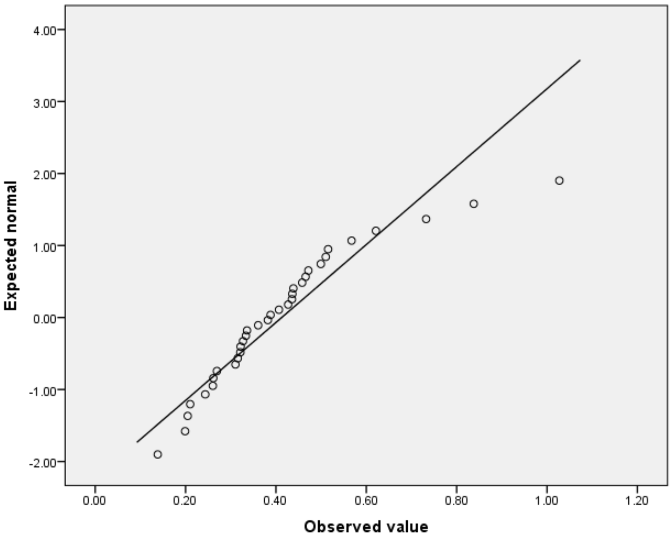


Figure 3. Normal Q-Q plot for the outlier data point (L2, ISI 1, Similarity 2).

In contrast, an outlier defined as an observation that does not approximately follow the regression line—as presented in Figure 4—might affect its slope, and subsequently, adversely affect the findings by increasing the chance of type I or type II errors. There were only two outliers meeting this consideration and these were excluded from the data analysis. The same reasoning was used across all analyses in the thesis.

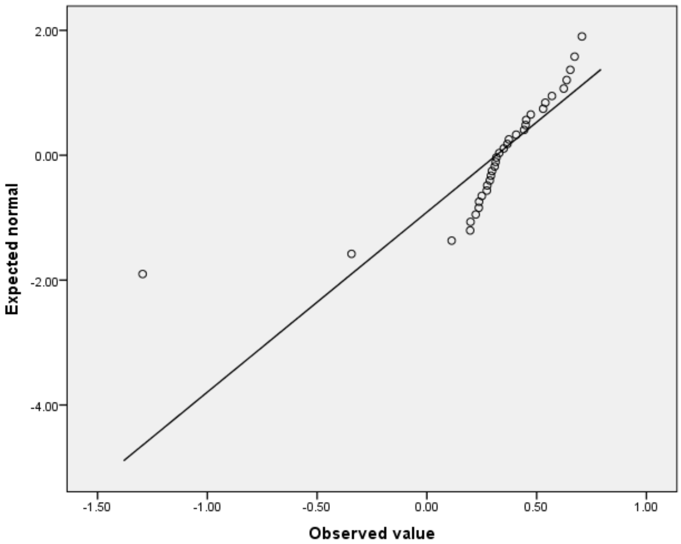


Figure 4. Normal Q-Q plot for the outlier data point (L2, ISI 3, Target).

The outliers could, however, inflate within-group variability. There has been debate around the robustness of multivariate analysis ANOVA (e.g., Field, 2012; Hoekstra, Kiers, & Johnson, 2012; Schmider et al., 2010) but, generally, it has been agreed that the assumption of normality matters, especially in smaller samples, such as the one in this experiment. The violation of assumptions can adversely affect the findings and their interpretation. For these reasons, the data were screened to verify that they met the assumptions required for the relevant statistical test to establish that the data met the requirements for the application of parametric statistical analysis. Where the assumptions were found to be violated, a non-parametric alternative was chosen.

Although Tabachnick and Fidell (2007) recommended transforming non-normally distributed data in an attempt to make it conform to normality, inspection of additional sources suggested that there is little consensus as to whether data transformation is better than using a non-parametric analysis. The choice likely depends on the research questions and the details of the design. (This will be discussed in further detail within the research questions and design of the current study.)

The main advantage of non-parametric analysis is that it is distribution-free. The data are ranked from lowest to highest and the analysis uses median instead of mean. In this study, some outliers were retained. If there were few participants who had very slow or very fast responses, the mathematical mean would increase or decrease greatly, but the RT of the typical participant would not change. Therefore, the median can arguably be a better measure of the central tendency of the data. The same reasoning was used across all analyses in the thesis. Finally, the level of statistical significance, alpha, was set at .05 for all statistical tests, and all tests were conducted as two-tailed.

Before conducting the analyses, two aspects of the obtained data were considered. First, the RT corresponded to a keypress made when an auditory stimulus was determined to be a target or non-target, and this may have occurred before the stimulus had finished. Second, it is possible that part of the variance in RT was caused by differences in the durations of spoken Chinese and English words. Therefore, the mean pure RT was calculated by subtracting the duration of the stimulus from the RT on that particular stimulus for each of the stimuli across all language conditions, and this was used as data for the analyses that followed. This meant that negative values of RT were possible. The mean durations of the stimuli and the mean pure RTs for each language condition are presented in Table 8.

Table 8

Mean Response Times (RT), Mean Durations of Stimuli, and Mean Pure RT in Seconds in First (L1), Second (L2), and Language Switching (Mix) Conditions

	Language conditions					
	L1	*SD**	L2	*SD*	Mix	*SD*
Mean RT	1.556	0.274	1.731	0.266	1.699	0.265
Mean stimuli duration (s)	1.252	0.083	1.549	0.082	1.399	0.214
Mean pure RT (s)	0.304	0.268	0.182	0.256	0.300	0.254

*SD = standard deviation

To provide a comprehensive analysis, and for clarity, the Results section is organised in line with two main indices of performance; speed and accuracy. Each of these sections is further organised according to the within- and between-subjects variables. Finally, SDT measures are reported.

5.3.2. Performance Speed

To explore performance differences between the three language conditions, a multivariate analysis was conducted. It is important to note that the primary focuses of the analysis were the main effects of the three factors (Language condition, ISI and Similarity) and interactions related only to the Language condition (Language condition and ISI factors; Language condition and Similarity factors). Evidence of a significant main effect would suggest a difference exists between the three language conditions (L1, L2 and Mix).

A 3 × 3 × 4 within-subjects ANOVA was performed to test whether there was a difference in speed of recognition attributable to the three different language conditions (L1, L2 and Mix), ISIs (1 s, 4 s, 9 s), and the degree of stimuli similarity (Similarity 0, Similarity 1, Similarity 2 and Target). As the repeated-measures ANOVA is sensitive to violations of sphericity, Mauchly's test was considered and was found to have been violated; therefore, degrees of freedom were corrected by Greenhouse-Geisser Epsilon (G–G ε) following Girden's (1992) recommendation. The same reasoning was applied throughout the analyses of other studies.

There was evidence of statistically significant main effects for the Language condition ($F(2, 66) = 20.162$, $p < .001$, $\eta_p^2 = .379$), ISI (Greenhouse-Geisser adjusted $F(1.25, 41.25) = 8.232$, $p = .004$, $\eta_p^2 = .200$), and for Similarity (Greenhouse-Geisser adjusted $F(1.44, 47.52) = 80.596$, $p < .001$, $\eta_p^2 = .709$). Two-way interaction effects were also found to be significant between the Language condition and ISI ($F(4, 132) = 4.178$, $p = .003$, $\eta_p^2 = .112$), between the Language condition and Similarity (Huynh-Feldt adjusted $F(5.5, 181.38) = 9.144$, $p < .001$, $\eta_p^2 = .217$), and between ISI and Similarity (Greenhouse-Geisser adjusted $F(3.78, 124.7) = 3.767$, $p = .007$, $\eta_p^2 = .102$). Finally, the three-way interaction between the Language condition, ISI and Similarity was statistically significant (Greenhouse-Geisser adjusted $F(5.06, 167.04) = 2.867$, $p = .016$, $\eta_p^2 = .080$). To determine exactly where the differences lay, attention was concentrated upon two-way interactions. The interaction plots are presented in Figure 5.

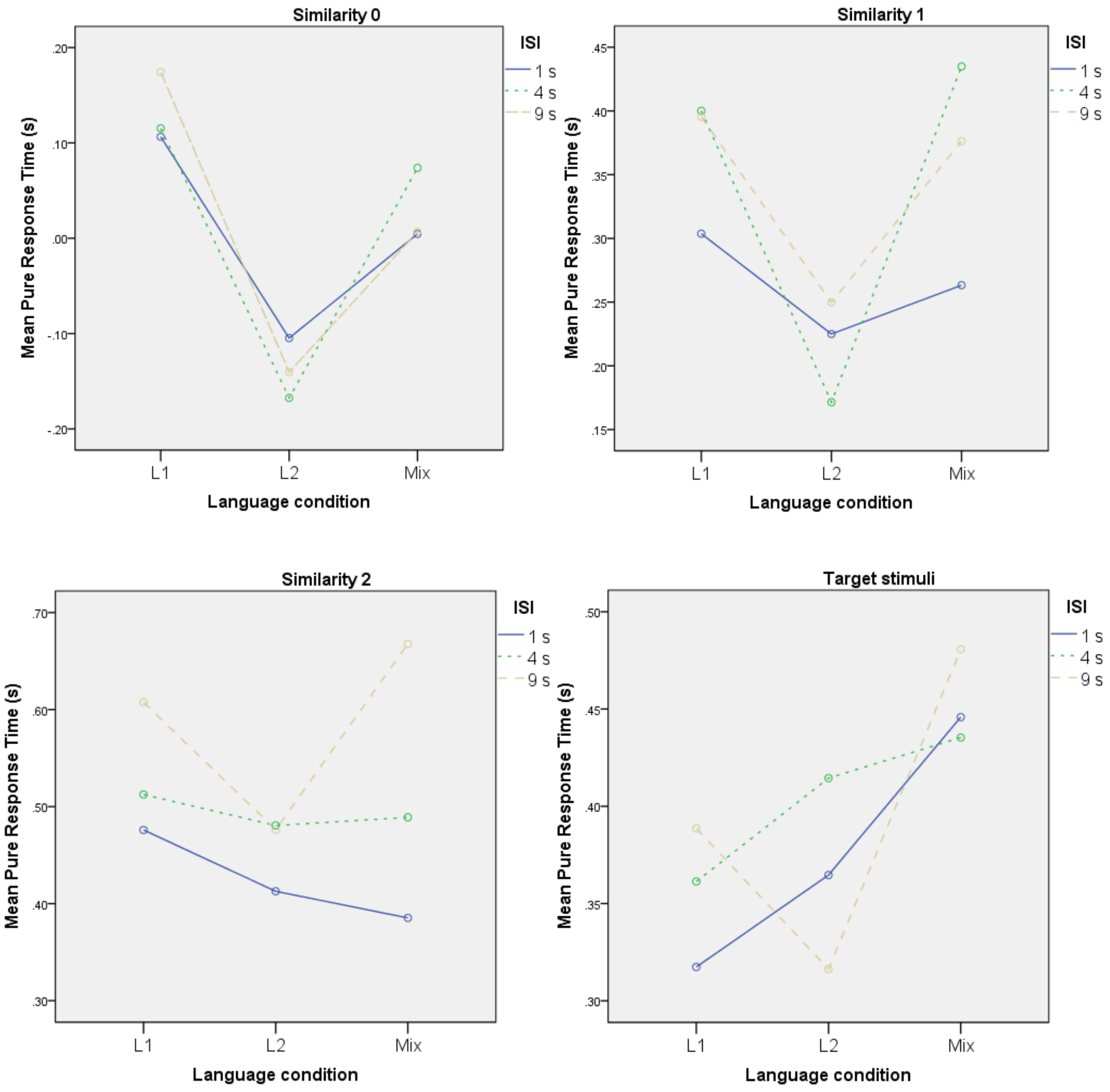


Figure 5. Interaction plots for Language condition, ISI, and Similarity factors.

As reporting *all* interactions would reduce the clarity of findings in this complex design, the *post hoc* analysis was conducted to determine only interactions related to the main effect of the Language condition and the related and predicted two-way interactions.

5.3.2.1. Performance Speed and Language Conditions

Prior to conducting the *post hoc* analysis, the assumption of normality was tested and was considered to be violated as the skewness and kurtosis levels of some of the data were beyond the span of –2.0 and +2.0 (Cramer, 1998; George & Mallery, 2010). The values of skewness and kurtosis are not presented in a table because of the large number of combinations, which would reduce the readability of the overall results. It was therefore decided that a non-parametric alternative was more appropriate for the analysis. No adjustment of alpha was made for the data under evaluation, as recommended by Rothman (1990).

A Wilcoxon signed-rank test was used for the *post hoc* analysis of the predicted main effect of the Language condition. Findings indicated that the performance was significantly faster in the L2 condition than in the Mix condition (*Mdn* = 0.161 vs. *Mdn* = 0.287; *Z* = –4.132, *p* < .001, *r* = .709), or the L1 condition (*Mdn* = 0.161 vs. *Mdn* = 0.282; *Z* = –3.922, *p* < .001, *r* = .673). However, no difference was found between the L1 and Mix conditions (*p* = .662).

5.3.2.2. Performance Speed and ISI

Given the violated assumption of normality, as the skewness and kurtosis levels of some of the data were beyond the span of –2.0 and +2.0 (Cramer, 1998; George & Mallery, 2010), a non-parametric Wilcoxon signed-rank test was conducted to test the differences across the levels of ISI. The findings indicated that in the L1 condition, the higher the interval between the stimuli, the slower the pure RT; difference between ISI 1s and 4 s (*Z* = –2.983, *p* = .003, *r* = .17), and between ISI 4 s and 9 s (*Z* = –3.821, *p* = .026, *r* = .13). There was no difference in speed of performance across all levels of ISI in the L2 condition (*p* ≥ .768). In the Mix condition, the performance was statistically significantly slower only on ISI 4 s in comparison to ISI 1 s (*Mdn* = 0.331 vs. *Mdn* = 0.265; *Z* = –4.231, *p* < .001, *r* = .24). Finally, performance in the L2 condition was significantly faster than in the Mix condition across all levels of ISI factor (*p* ≤ .048; –.79 ≤ *r* ≤ –.34). The negative sign of the effect sizes, *r,* reported the direction of the effect. This was based on the order of the sample means. Table 9 summarises the data for the two-way interaction of Language and ISI factors.

Table 9

Median (Mdn) Pure Response Times in Seconds across Language Condition and Inter-stimuli Interval (ISI) Factors

	ISI 1 s	ISI 4 s	ISI 9 s
	Mdn	*Mdn*	*Mdn*
L1	0.257	0.299	0.343
L2	0.143	0.157	0.165
Mix	0.265	0.331	0.315

5.3.2.3. Performance Speed and Similarity

Prior to testing the effect of different levels of similarity on speed of performance, the assumption of normality was tested and was considered to be violated as the skewness and kurtosis levels of some of the data were beyond the span of –2.0 and +2.0 (Cramer, 1998; George & Mallery, 2010). Therefore, a Wilcoxon signed-rank test was conducted, which indicated that the higher the level of similarity, the slower the pure RT in all language conditions ($p \leq .002$; $-.86 \leq r \leq -.54$), except the performance on Similarity 2, which was slower than on the Target number in all language conditions ($p \leq .050$; $-.86 \leq r \leq -.34$). The performance was statistically significantly faster in the L2 than in the Mix condition across all levels of the Similarity factor ($p \leq .014$; $-.67 \leq r \leq -.46$), except Similarity 2, where no difference was found ($p = .161$) between the two language conditions. Medians are presented in Table 10.

Table 10

Median (Mdn) Pure Response Times in Seconds across Language Condition and Similarity Factors

	Similarity 0	Similarity 1	Similarity 2	Target
	Mdn	*Mdn*	*Mdn*	*Mdn*
L1	–0.007	0.296	0.467	0.311
L2	–0.244	0.134	0.387	0.351
Mix	–0.094	0.324	0.475	0.442

5.3.2.4. Switch Costs and Mixing Costs

To test the hypothesis of the presence of asymmetric switch and mixing costs, further analysis of the sequence of alternating English and Chinese stimuli in the Mix condition was conducted. Prior to testing whether the differences were statistically significant, the assumption of normally distributed differences between pairs was examined and was found to be violated as the skewness and kurtosis levels were beyond the span of –2.0 and +2.0 (Cramer, 1998; George & Mallery, 2010). Therefore, a non-parametric Wilcoxon signed-rank test was conducted, which indicated the presence of asymmetric mixing costs. In other words, the performance was slower when switching to participants' native language after hearing a stimulus in participants' second language (*Mdn* = 0.41), than other way around (*Mdn* = 0.28) in the Mix condition (*Z* = –2.94, *p* = .003, *r* = .25). Data are presented in Table 11.

Table 11

Median (Mdn) Pure Response Times in Seconds on Chinese and English Word Stimuli in Monolingual (L1 and L2) and Language Switching (Mix) Conditions: Mixing Costs and Switch Costs

Monolingual		Mix			Switch costs
Sequence	*Mdn*	Sequence	*Mdn*	Mixing costs	(Pure vs. Mix)
L1 → L1	0.28	L2 → L1	0.41	–0.13	→ L1: 0.13
L2 → L2	0.15	L1 → L2	0.28		→ L2: 0.13

There was evidence of statistically significant switch costs as well. The responses on Chinese stimuli in the L1 condition were significantly faster than responses on Chinese stimuli in the Mix condition (*Z* = –2.37, *p* = .018, *r* = .20). Similarly, the responses on English stimuli in the L2 condition were faster than those in the Mix condition (*Z* = –2.95, *p* = .003, *r* = .25). However, there was no difference between these switch costs; that is, the difference in RTs between the Chinese stimuli on the L1 and the Mix conditions was of similar duration to the difference in RTs between the English stimuli on the L2 and the Mix conditions, (*p* = .724).

5.3.2.5. Effect of English Language Proficiency on Performance Speed

Prior to investigating the effect of participants' English language proficiency on performance speed, the assumption of normality for a between-subjects ANOVA was evaluated and was found to be satisfied as the two groups' distributions were associated with skewness and kurtosis within the span of –2.0 and +2.0 (Cramer, 1998; George & Mallery, 2010). Next, the assumption of homogeneity of variances was tested and was found to be satisfied, based on Levene's *F* test, in the L2 condition ($F(1, 32) = 0.021$, $p = .886$), and in the Mix condition ($F(1, 32) = 1.783$, $p = .191$). Therefore, a between-subjects ANOVA was conducted; however, there was no evidence of a statistically significant difference in mean pure RTs between the two English language proficiency groups in either condition (L2, $p = .178$; Mix, $p = .261$). Data associated with participants' mean pure RT across the two English language proficiency groups are reported in Table 12 and graphically displayed in Figure 6.

Table 12

Mean Pure Response Times in Seconds across the English Language Proficiency Levels

	IELTS Listening	*n*	L2	*SD*	Mix	*SD*
Modest/Competent User	5.0–6.5	18	0.241	0.252	0.348	0.219
Good/Very Good User	7.0–8.5	16	0.116	0.275	0.244	0.307

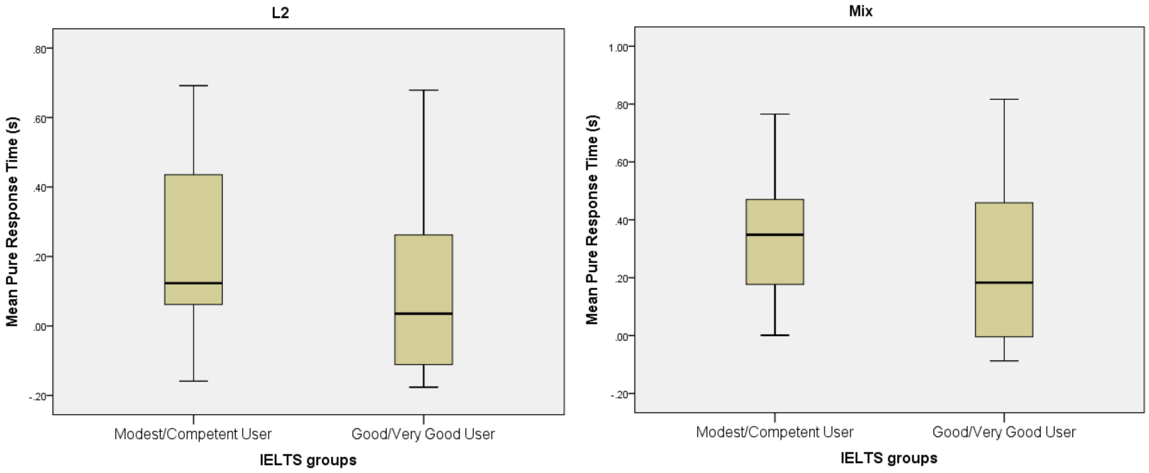


Figure 6. Profile plots for English language proficiency groups in the L2 and Mix conditions

.

5.3.3. Performance Accuracy

Although errors occurred in less than 1% of the language conditions, an error analysis was performed. There was sound justification for this: Because of the high number of flights occurring on average each day (more than 100,000 scheduled flights per day, according to the International Air Transport Association, 2014), even such small error rates may have important implications for flight safety. The analysis sought to explore the nature of the errors to minimize the potential risk by increasing awareness of potential hazards.

5.3.3.1. Performance Accuracy and Language Conditions

Prior to testing whether performance accuracy differed between language conditions, the assumption of normality was evaluated and was found to be violated as the distributions were associated with skewness and kurtosis beyond the span of –2.0 and +2.0 (Cramer, 1998; George & Mallery, 2010). Therefore, the non-parametric Friedman's test was used, which indicated no statistically significant difference in accuracy across the language conditions (*p* = .560). Data are presented in Table 13.

Table 13

Error Types (Miss and False Alarm), Hits and Correct Rejections (CR), and Total Number of Errors across Language Conditions, and Percentage of Errors from 5100 Stimuli (%Error$_T$)

	Miss	False alarm	Hits	CR	Error Total	%Error$_T$
L1	3	9	541	1147	12	
L2	14	3	530	1153	17	
Mix	9	5	535	1151	14	
Total	26	17	1606	3451	43	.843

5.3.3.2. Performance Accuracy and ISI

Prior to testing whether performance accuracy differed between the ISIs, the assumption of normality was evaluated and was found to be violated as the distributions were associated with skewness and kurtosis beyond the span of –2.0 and +2.0 (Cramer, 1998; George & Mallery, 2010). Friedman's test indicated statistically significant differences ($\chi^2(2) = 36.033$, $p < .001$). The error counts and rates across the three ISIs are presented in Table 14. Wilcoxon signed-rank tests further indicated that the error rates increased with increased ISI, between ISI 1 s and 4 s (0.58% vs. 0.81%, $Z = –2.000$, $p = .046$, $r = .34$, risk ratio = 0.714), and ISI 4 s and 9 s (0.81% vs. 1.16%, $Z = –4.021$, $p < .001$, $r = .68$, risk ratio = 1.442).

Table 14

Number (n_{errors}) and Percentage of Errors (%Error$_E$) across ISI Levels and as a Proportion of Total Number of Stimuli ($n_{stimuli}$) in Each ISI Level (%Error$_{ISI}$)

	ISI 1 s	ISI 4 s	ISI 9 s
n_{errors}	10	14	19
%Error$_E$	0.58	0.81	1.16
%Error$_{ISI}$	0.20	0.27	0.37
$n_{stimuli}$	1734	1734	1632

5.3.3.3. Performance Accuracy and Similarity

Owing to the violated assumption of normality, based on the distributions associated with skewness and kurtosis beyond the span of –2.0 and +2.0 (Cramer, 1998; George & Mallery, 2010), a non-parametric Friedman's test was used, which indicated significant differences in accuracy across the Similarity levels ($\chi^2(3) = 35.166$, $p < .001$). Wilcoxon signed-rank tests indicated no statistically significant difference between Similarity 0 and 1 ($p = .679$). The risk of an error on Similarity 2 was almost six times the risk of an error on Similarity 1 ($Z = –3.162$, $p = .002$, $r = .54$, risk ratio = 5.999), and almost nine times the risk of an error on Similarity 0 ($Z = –3.292$, $p = .001$, $r = .56$, risk ratio = 8.999). No difference in accuracy was found between Target and Similarity 2 ($p = .065$). The risk of an error on

Target was approximately six times the risk on Similarity 1 ($Z = –3.504$, $p < .001$, $r = .600$, risk ratio = 6.499), and almost ten times the risk on Similarity 0 ($Z = –4.945$, $p < .001$, $r = .848$, risk ratio = 9.749). Even though risk ratio also refers to the strength of the association between the variables, as effect sizes do, its additional value was that it provided a concrete percentage of the risk of errors (Schmidt & Kohlmann, 2008). Risk ratio was chosen instead of odds ratio because of the low occurrence of errors. In situations of small error rates, the values of odds and risk ratios would be very similar anyway (Schmidt & Kohlmann, 2008). Data are presented in Table 15.

Table 15

Number (n_{errors}) and Percentage of Errors (%Error$_E$) across Levels of Similarity Factor (%Error$_S$), and as a Proportion of Total Number of Stimuli ($n_{stimuli}$) in Each Level of Similarity Factor ($n_{stimuli}$)

	Similarity 0	Similarity 1	Similarity 2	Target
n_{errors}	3	2	12	26
%Error$_E$	7	4.7	27.9	60.5
%Error$_S$	0.16	0.25	1.47	1.59
$n_{stimuli}$	1836	816	816	1632

5.3.3.4. Effect of English Language Proficiency on Performance Accuracy

The assumption of normality was evaluated for the English language proficiency data and found to be violated, based on the distributions associated with skewness and kurtosis beyond the span of –2.0 and +2.0 (Cramer, 1998; George & Mallery, 2010). Therefore, the relation between error occurrence and English language proficiency was tested using a non-parametric Mann-Whitney test. However, the test indicated no difference in performance accuracy between the groups of English language proficiency ($p = .296$). The number of errors and correct responses across English language proficiency groups are presented in Table 16.

Table 16

Frequency of Correct Responses and Errors, and Percentage of Errors (%Errors) across English Language Proficiency Levels

	n	Correct	Errors	%Errors
Modest/Competent User	18	2680	20	0.75
Good/Very Good User	16	2377	23	0.97

5.3.4. SDT Measures

The total counts of hits, misses, correct rejections and false alarms across all three language conditions were calculated, from which the hit rate (HR; number of hits/number of target stimuli) and the false alarm rate (FAR; number of false alarms/number of distracting stimuli) were derived. The rates were used to test for response bias by calculating sensitivity, d' = z(HR)–z(FAR) and decision criterion C = –0.5[z(HR) + z(FAR)] in accordance with Stanislaw and Todorov’s (1999) SDT calculations. Hit and false alarm rates by language condition are summarised in Table 17, along with respective values of *d’* and *C*.

Table 17

Sensitivity (d') and Decision Criterion (C) of Call Sign Recognition Task across the Three Language Conditions: Native (Chinese) Language (L1), Second (English) Language (L2), and Language Switching (Mix)

	z(HR)	*z(FAR)*	*d'*	*C*
L1	2.54	–2.42	4.96	–0.061
L2	1.95	–2.80	4.75	0.425
Mix	2.13	–2.56	4.69	0.215

The values of *d’* and *C* suggest that participants had little difficulty in distinguishing target from distracting stimuli in all language conditions ($d' \geq 4.69$). There was also evidence of a bias towards a *no* response—that is, a tendency to miss the target in both the Mix and the L2 conditions ($C \geq 0.215$)—and a small response bias towards a *yes* response in the L1 condition ($C = –0.061$).

5.4. Discussion

The findings indicated that recognition performance in the L2 condition was faster than in the Mix condition. However, there was no difference in performance accuracy between the two conditions. The longer stimuli RTs in the Mix condition may suggest greater cognitive demands when identifying acoustic speech stimuli in a bilingual than a monolingual environment. This finding is consistent with previous studies (Bobb & Wodniecka, 2013; Costa & Santesteban, 2004a; Meuter & Allport, 1999). Interestingly, however, the size of the switch costs appeared to be the same (see Table 11). This suggests that it may be the *alternation* between languages that creates the additional processing costs rather than the different cognitive demands of dominant and non-dominant language.

It may appear counter-intuitive that performance in the participants' second language was faster than in their first language. One explanation may be that participants, at the time of participating in the experiment, had been living in an English-speaking country, and thus communicated in English on a daily basis. Additionally, the task itself did not put large demands on language skills, and it might therefore have been easier to respond in the language that was activated most of the time.

Because of the chosen approach to RT measurement, a comparison between mean RT and mean pure RT is also made. It is speculated that the analysis could have resulted in opposite findings if the mean RT, instead of pure RT, had been used for the analysis, with performance speed in the L2 condition (1.731 s) being slower than in the L1 (1.556 s) and Mix (1.699 s) conditions (see Table 8). Even though this would not confirm the correctness of the approach to RT measurement that was used, it could indicate that there is an effect of the different stimuli duration on performance speed. However, it is unclear whether the subtraction of the stimuli length from the mean RT would always yield shorter RTs in the L2 condition, because of the generally longer duration of English words in comparison with Chinese. A question that might serve to verify the reliability of this method is whether even longer latencies might be found in the L2 condition. Further investigation would be necessary to provide a reliable answer to this question. At this stage, the observed difference can only be attributed to the rationale discussed in section 3.12; that is, the cognitive processing time of language should be independent from the duration of the stimuli.

Although a difference in speed of performance was found between different ISIs in the L1 condition, it was not found in the L2 condition. In the Mix condition, the difference occurred only between the short and medium intervals (1 s vs. 4 s), but not with the longer interval (9 s). The findings also suggested that the greater the interval between the stimuli, the greater the number of errors. Although no research question or hypothesis was made regarding the ISI, this pattern of findings was somewhat unexpected; as such, replication and further investigation would be needed to discover if this finding is robust and why it occurred.

According to the findings, there was no evidence that the observed differences were attributable to the level of English language proficiency. However, as previously discussed, the task was very simple and did not place significant demands on language skills. Proficiency in the second language may be a more important factor for the processing of more complex sentences (Barshi & Farris, 2013).

There was some evidence that the higher the level of Similarity, the slower the RT in all language conditions, except the performance on Similarity 2, which was slower than the performance on target number in all language conditions. Negative values of RT on Similarity 0 (see Table 10) were interpreted in accordance with Marslen-Wilson's (1985) finding, that the stimuli are processed before they are heard completely. For example, when the RT was 1.9 s and the duration of the word was 1.6 s, then the pure RT was 0.3 s. In this case, participants pressed "yes" or "no" when they heard the whole number. However, when the Similarity of the target and distracting numbers was 0, participants tended to respond while the stimulus was playing. That is, if the duration of the stimulus was 1.6 s and the RT was 1.36 s, it meant that a participant pressed a response key within 1.36 s of the word being spoken and .24 s prior to its finish. This finding may provide evidence to support recognition primed decision making (RPDM), as the RTs increased with increased Similarity. This intuitive decision-making model is based on recognising patterns in a new situation (currently playing number) that match patterns stored in memory (known target number) (Jensen, Guilke, & Tigner, 2005).

An explanation of the findings of performance speed using the RPDM might be as follows. When the target was "531" and participants heard a number that began with "7_ _" (Similarity 0), they immediately recognised the difference and decided to press "no." The same strategy was applied for Similarity 1 across all language conditions. Interestingly, however, there were longer latencies for the stimuli with Similarity 2 than for Targets. It is possible that participants were able to recognise the Target number more quickly because it was stored in memory and was more familiar than an unknown new number.

The findings related to performance speed are broadly in accordance with Cox and Vinagre's (2004) study, where the digit in the middle of a three-digit phrase was more likely to be misrecognised than either of the outer digits. This observation could be explained also within the serial-position effect (VandenBos, 2015), which was identified in studies on memory (recalling words from a list), although not yet explored within studies of recognition (recognizing new items). The effect suggests that the position of an item in a list of items to be learned affects how well it is remembered. It was found that the first and the last items in a series are best recalled while the middle items are worst (VandenBos, 2015). Future research may explore as to whether the position factor also affects the recognition of numeric call signs. Corradini and Cacciari (2002) debated that carrier companies usually attribute call signs to their aircraft that differ only by the last letter when aircraft follow the same route (e.g., KLM254a and KLM254b). In the context of these findings it is thought-provoking as to what would be the difference in performance when these call signs differed in the first number instead (e.g., KLM654 and KLM254), or if the last letter was shifted between "KLM" and the number "254" (e.g., KLMa254 and KLMb254). Based on Cox and Vinagre's (2004) findings, there would likely be more mistakes when the middle digits differed (e.g., KLM254 and KLM234). Corradini and Cacciari (2002) noted that ATCOs often shorten call signs, eliminating either the company name or the numerical string, which makes read-backs less effective in disambiguating the miscommunication.

Although RPDM has not previously been applied to the call sign recognition experiment, support for this interpretation might be found in previous studies of SA and RPDM (Klein, 2000). Even Endsley's conceptualisation of SA (1995, 2000) suggests that the acquisition of SA is achieved through a process of pattern-matching with previous experience. This can then be used as an explanation for the finding from Study 1, that participants may not pay attention to communications that are in a language they do not understand (*"As pilots we almost all have a tendency to "tune out" non-essential transmissions,"* Participant 76, pilot). Consequently, SA can be impaired, which can, in turn, critically affect aviation safety.

The impact of call sign similarity on recognition performance that would affect safety was more transparent when performance accuracy was considered. It was found that the higher the level of similarity between the distracting stimuli and target, the higher the number of errors. The risk of missing a target stimulus was 10 times greater than of making an error on Similarity 0. In aviation, a miss type of error may potentially have more serious consequences than a false alarm. The response bias toward non-detection can potentially be attributed to a higher occurrence of distracting stimuli than infrequent targets. This may create a higher expectation of distracting stimuli, and, consequently, participant response bias. However, in the case of Similarity 2, when only the last digit differed from the target, the risk of making a false alarm error did not differ from the risk of missing a target.

It must be also noted, that the two errors made on Similarity 1 contained reversed digits (e.g., the target was "729" and the number of Similarity 1 was "792") causing a certain similarity with the target number, which might have been the primary cause of the false alarm responses. This may suggest that the position of the digits in a numerical call sign is a significant factor. This would be consistent with the results of the ACCESS study (CAA, 2000), which found that 13% of confusion occurrences involved the use of the same digits in a different order. Interestingly, however, in the ACCESS study (CAA, 2000) the Similarity factor was considered starting from the end of the call signs, rather than beginning as was done in this study. Yet, the number of errors also increased with an increased number of same digits. Therefore, it can be debated whether the confusion is caused by the number of the identical digits (items) or their order in a call sign. More research would be needed to answer this question.

Situations in which participants listen for the minimum information required to make a decision on whether there is a signal can create opportunity for errors. Even though the overall occurrence of errors was less than 1% in all language conditions, which is common in language switching experiments (for reviews see Zheng, et al., 2018), it can be assumed that in real-life situations, the rate of errors may be higher. Considering the high number of flights occurring on average each day (more than 100,000 scheduled flights per day, according to the International Air Transport Association, 2014), even such small error rates may have important implications for flight safety.

There are three potential limitations of the current study. First, as discussed previously, the study involved the use of numeric call signs only; whether similar results would be obtained for alphanumeric call signs or three-letter designators for aircraft cannot be determined.

Second, caution must be applied prior to generalizing the current findings to aviation personnel because the participants in the current study were non-aviation students performing cognitive tasks, whereas pilots perform complex duties of flying, navigating and communicating. However, it should be noted that the potential problems raised by this limitation are not unique to aviation psychology studies, as a large proportion of all psychological studies have relied on generic students as participants (Wintre, North, & Sugar, 2001) Presumably, the effects found in this study influence the performance of most people in the same direction and therefore the primary variability is likely to be in terms of the degree, rather than the nature of the effect (Orlady & Orlady, 1999). This study captured the fundamental cognitive processing of language in general rather than specific realistic operations or complex behaviour.

Third, there are some concerns related to the self-reported IELTS test scores. The scores may not reflect the actual English language proficiency of the participants because the subsequent length of their stay in an English-speaking country may have affected their language competence. Moreover, no evidence of participants' IELTS test scores was required to confirm the reliability of the self-reported proficiency. However, with regard to both points, it is noted that using a self-reported proficiency of English is common in language switching studies (e.g., Koch, Decklerck, & Philipp, 2015; Zheng, Roelofs, & Lemhofer, 2018). The assumptions about the English language proficiency of aviation personnel, who must meet the ICAO English Language Proficiency Requirements expressed in the ICAO Rating Scale (2011), were made based on the findings of the IELTS test. Needless to say, the alignment of these two tests is not precise. Although this approach is common in aviation studies (e.g., Barshi & Farris, 2013), no precise conclusion about the English proficiency of aviation personnel can be drawn from the findings of this study.

In conclusion, the findings of the current study suggest that language conditions had some effects on the speed and accuracy of performance in this study group. These findings also suggest that the similarity of the stimuli and ISI may affect acoustic speech recognition performance.

CHAPTER SIX

Study 3: Error Identification

6.1. Introduction

To be situationally aware, pilots and air traffic controllers (ATCOs) maintain listening watch on the control frequency. Maintaining listening watch in a bilingual air traffic environment can be more difficult, given that the additional task of language alternation is required. Safe and efficient air traffic control (ATC) is reliant upon accurate communication between pilots and ATCOs, where any communication error has the potential for a significant consequence (Auton et al., 2016). The ability to detect such errors can be crucial for safe operations. Presumably, routine situations can cause task performance to become automated and, thus, decrease error detection. It was the aim of the second Bilingual IFR Communications Simulation Study (BICSS, Borins, 1983) to investigate listening watch in terminal operations.

In the BICSS ATCOs were asked to make errors that led to a loss of separation between aircraft, to investigate whether pilots were able to identify erroneous clearances (Borins, 1983). This seems to be a common procedure when studying the second level of SA, as detecting an error in an ATC message requires its comprehension. Pew (2000) suggested the typical measure of performance to be the time required to detect an anomaly. This is important beyond pilot–controller communication.

To explore error identification, simple arithmetic equations were developed. The first study explored the recognition of three-digit numbers, and it was therefore decided to continue using numbers. This would allow the issues of the experimental research methodology proposed in section 3.8 to be addressed. The identification of mistakes in arithmetic equations represents a simple task exploring whether participants will notice that something is different to the way it should be. Applying this rationale, arithmetic problem solving may represent problem solving in general, which is a crucial aspect of aviation safety. Arithmetic operations may occur in pilots' operations with numbers (e.g., altitude, flight level, speed, course and call signs). Operations using numbers were found to be one of the most repeated factors in aircraft accidents or incidents (Tiewtrakul & Fletcher, 2010). Therefore, the aim was to investigate error detection in simple arithmetic problems (e.g., 2 + 4 = 8) in monolingual and bilingual conditions. The two guiding research questions for Study 3 were:

Question 3a: Do speed and accuracy of error identification differ between monolingual and bilingual conditions?

Question 3b: Is there a difference in performance speed and accuracy when responding to correct information and when identifying a mistake?

6.1.1. Detection of Erroneous Message

Based on a content analysis performed on 50 hours of pilot and ATC messages that were transmitted from five of the busiest terminal radar approach control facilities in the United States between October 2003 and February 2004, Prinzo, Hendrix, and Hendrix (2006) found that ATCOs corrected only 8% of the pilots' read-back errors. Of the corrected read-backs, almost 14% involved omission (e.g., omission of number element or anchor word; "two thousand" or "twelve" instead of "climb and maintain one two thousand"), 79% involved substitution (e.g., substitution of message numbers; "three one zero" instead of "two one zero"), and almost 7% involved transposition errors (e.g., transposition of message numbers; "climb two one thousand" instead of "climb and maintain one two thousand"). The findings showed that only 2% of all the omission errors (8/61), 19% of the substitution errors (46/243), and 21% of the transposition errors (4/19) were corrected. Prinzo et al. (2006) were interested in why were so few errors corrected.

They attributed the small proportion of erroneous read-back corrections to the insignificance of some errors, which may not have warranted another transmission, and they assumed that ATCOs would otherwise intervene when it would be necessary to maintain safety. The authors appear to have meant that the ATCOs detected but did not correct all the errors. However, this assumption cannot be confirmed, as the non-correction of an error may also indicate its non-detection. All the read-backs were uttered in English, and it is not yet known whether the situation would differ in a bilingual air traffic environment; there do not appear to be any similar analyses of erroneous read-back detection/correction in bilingual operations. Venkatraman et al. (2006) studied the effect of language switching on arithmetic and assessed the role of language in forming arithmetic representations; that is, solving arithmetic problems in different languages. Importantly, the participants were Chinese–English bilinguals, as in the current study. While solving the same arithmetic problems in both languages, they were scanned using Functional Magnetic Resonance Imaging (fMRI). It was found that the processing relied on verbal and language-related networks.

According to Wang et al. (2007), calculation in second language involves additional neural activation, especially in the left hemisphere, including Broca’s area. They concluded that the interaction between language and mathematics involves a specific neurocircuitry when associated with the second language. Van Rinsveld et al. (2016) studied how bilinguals' performance in arithmetic might be changed by setting the problems in a language context. The task in a language context was structured by presenting by a sentence in the same language as the arithmetic problem prior to the arithmetic problem itself (analogous to monolingual context). After participants made a semantic judgment on the priming sentence, they solved the arithmetic task in the same language. Van Rinsveld et al. (2016) compared this context condition to a no context condition where participants had to solve only arithmetic problems in the instructed language without any context. They found that providing a language context enhanced the arithmetic performance in the bilinguals’ second language. This raises the question of whether monolingual air traffic environment can provide a context that would facilitate faster and more accurate performance than bilingual air traffic environment.

To compare performance between language conditions, two studies (McClain & Huang, 1982; Marsh & Maki, 1976) explored performance in solving arithmetic problems in the preferred language (the language in which participants had first learned arithmetic) and non-preferred language. The findings suggested two possible interpretations, which can be consistent with the triple code model described in section 3.7. Typically, faster performance was observed in the preferred language than in the non-preferred (McClain & Huang, 1982; Marsh & Maki, 1976). Participants may have translated the problems (e.g., 2 + 3) into the preferred language to carry out the calculations. In this interpretation, RT was slower in the non-preferred language because of the translation process (i.e., language-specific format). The second interpretation was that participants translated the problems into abstract language-independent representations, and then translated the answers into the preferred or non-preferred language for output. Then, the RT was slower in the non-preferred language because of the translation of answers from abstract to language-specific form (format-independent account of arithmetic fact retrieval).

Geary et al.’s (1993) supports the claim that solving arithmetic problems when switching between the preferred and non-preferred languages in a single session, slows overall solution times. Bilinguals seemed to solve arithmetic problems more slowly than monolinguals, and this difference in problem-solving speed tended to increase with increased complicacy of the problems (Geary et al., 1993). McClain and Huang (1982) found that when Chinese–English bilinguals performed tasks in a monolingual condition, either using their preferred or non-preferred language, the RTs in the two languages were equivalent and the advantage of preferred language was eliminated.

Although McClain and Huang (1982) chose the auditory presentation of stimuli, which is consistent with this study, they also sought verbal answers to the arithmetic problems, which is in contrast. In other words, McClain and Huang’s (1982) participants’ task was to compute and say the solution of a presented arithmetic problem. Neither the study by McClain and Huang (1982), nor the one by Marsh and Maki (1976) analysed error identification performance using the equation verification method—that is, judging equations as true or false (Reder, 1982)—and therefore, the findings provide only limited value for the current study. Geary et al. (1993) used the equation verification method of both correct and incorrect mathematical equations and, therefore, the findings are more explanatory, because there appears to be no literature focusing on mental arithmetic within traditional language switching paradigms. Consequently, little is known about the impact of language alternation on mathematical problem solving and error identification.

Finally, an important remark that considers the odd–even rule (e.g., Krueger, 1986) is put forward, given its effect in sum verification experiments. The odd–even rule refers to the division of numbers into odd and even (Shepard, Kilpatric, & Cunningham, 1975). Thus, when verifying, for example, the sum ‘2 + 2 = 5’, it is obviously false without a need to calculate or retrieve the correct sum. The odd–even property of numbers is salient. Krueger (1986) stated that when two numbers are added, the true sum must be even if both addends are even (e.g., 2 + 4, 8 + 6, etc.) or if both are odd (e.g., 3 + 7, 1 + 9, etc.); otherwise, it must be odd (e.g., 1 + 2, 3 + 6, etc.). The odd–even rule can be a cue that may permit participants to bypass normal processing, such as calculation or retrieval of correct response (Krueger, 1986). As such, it can facilitate error identification in arithmetic problems, and instead of coming to a correct result, it can provide a quick estimation about the correctness of an information. This indicates a rather passive cognitive process. However, in aviation, estimations of message correctness may not necessarily be sufficient.

With respect to the odd–even rule, this type of task (i.e., equation verification) was considered to meet the criteria for the thorough analysis, given the primary objective of comparing the performance of error identification in different language conditions.

6.2. Method

6.2.1. Overview

A computer-based experiment was developed in which participants were required to identify an error in simple arithmetic equations in either English, Chinese, or a Mix of both languages. Participants were Chinese–English bilinguals, following the same rationale as in Study 2.

6.2.2. Participants

A total of 40 non-aviation Chinese–English bilingual students (20 female and 20 male) participated in this study. The mean age of the participants was 28.63 years (*SD* = 8.75; *Range* = 18–53 years). All participants were enrolled at Massey University, New Zealand. English language was the participants' second language and Chinese, Mandarin dialect, was their native language. The mean duration of their stay in New Zealand was 2.31 years (*SD* = 3.25; *Range* = 1 month–19 years). Participants completed the IELTS test as part of the experimental session and were assigned to one of four English language proficiency groups (from Modest to Very Good) according to the score they achieved. No participant reported known hearing impairment.

6.2.3. Design

6.2.3.1. Overview

The primary aim of this study was to test the effect of language switching on error identification. A 3 × 3 within-subjects experimental design was used. The two within-subjects factors were Language condition (Chinese, English or a Mix of both languages) and ISI (the time interval between one stimulus and the next, which was 1, 4 or 9 seconds). A secondary aim was to test the effect of the between-subjects factor of language proficiency (ranging from Modest to Very Good) on performance in the English language and Mix conditions.

The benefits and potential shortcomings of using these designs were discussed in Study 2, namely the opportunity to manipulate more independent variables simultaneously and observe their effects on the dependent variable (within-subjects design), and the opportunity to compare performances regarding the dependent variable (between-subjects design). This methodology is commonly used in aviation psychology research (e.g., Barshi & Farris, 2013; Estival, Farris & Molesworth, 2016).

6.2.3.2. Development of the Experiment

The speed and accuracy of error identification were explored in three different language conditions: pure Chinese language condition (L1), pure English language (L2), and the language switching condition (Mix), composed of both English and Chinese stimuli. The presentation order of the language conditions was randomized. However, to avoid performance differences caused by the persisting inhibition of irrelevant language from the preceding condition (e.g., Meuter & Allport, 1999), the two monolingual conditions (L1 or L2) were not presented consecutively.

In each language condition, the participants' task was to identify incorrect and correct mathematical equations. Each language condition was developed to satisfy the following *criteria*: (i) only simple equations were chosen, requiring calculations up to 20, to ensure that the performance would not be adversely affected by the difficulty of solving a mathematical problem, and the findings would be attributable to the cognitive process of comprehension; (ii) the number of additions and subtractions within the stimuli list and also within the correct and incorrect equations were balanced; (iii) the equations were different across all language conditions; and, (iv) the difficulty of the conditions was equivalent, to ensure that the results were attributable to language conditions rather than the difficulty of either of the conditions. Despite these four criteria, the cognitive process by which participants come to a decision about correctness of the equations will remain opaque to the researcher; that is, it is possible that participants might just guess the answers instead of attempting to solve them by reasoning. Consequently, guessing might decrease the actual response times. An attempt to control for the potential confound of guessing was made in the test instructions, which stressed both accuracy and speed of responses. Stressing the accuracy of responses may control for the random guessing, provided that participants would be motivated to make correct judgements.

The experiment was created using the open-source application PsychoPy 1.82.01. (Peirce, 2007), in which the order of presentation of the stimuli in each language condition was set randomly to minimize the potential for any learning effects. The experiment was designed using the guidelines provided by the Massey University Ethics Committee and was deemed to be of low risk by a peer review. A copy of the institutional low-risk notification can be found in Appendix D.

6.2.3.3. Measures

The dependent variables were RT and the type and number of errors. Two types of errors were analysed, misses (e.g., participants responded to an incorrect equation as if it was correct) and false alarms (e.g., participants responded to a correct equation as if it was incorrect).

Two within-subjects independent variables were investigated, Language condition and ISI. The Language condition factor had three levels; equations were presented either in Chinese (L1) only, in English (L2) only, or in the language switching condition (Mix) of alternating Chinese and English stimuli. In each of the language conditions, the ISIs of 1 s, 4 s or 9 s were randomly distributed. ISIs of 1 s appeared 14 times per monolingual language condition (L1 and L2) and 17 times in the Mix condition. The ISI of 4 s and 9 s each appeared 13 times per monolingual language condition (L1 and L2). In the Mix condition, the ISI 4 s was randomly distributed 17 times and the ISI 9 s appeared 16 times.

Participants' ability in English language was a between-subjects independent variable. It was measured by the IELTS Listening scores they achieved when they completed the IELTS test as part of the study procedure.

6.2.4. Materials

6.2.4.1. Acoustic Stimuli

To test the effect of language conditions on the identification of infrequent errors in messages, simple arithmetic problems (additions and subtractions) were chosen as the acoustic stimuli. All stimuli were spoken by the same computerised female voices (using OS X Text-to-speech programme) in Chinese and English language as in Study 2, and at the same speech rate. Although ICAO (2001a) recommended a rate of 100 words per minute for aviation based radio communication, the decision was made to set the speech rate to 140 words per minute. The rationale for this was to simulate natural speech rate that participants (i.e., non-aviators) would be familiar with. Because the equations consisted of only 5 words, the speech rate of 100 words per minute appeared distractingly and pedantically slow. It was decided to record Chinese language stimuli in only the Mandarin dialect for two reasons. First, the occurrence of the Cantonese dialect in Study 2 had been minimal and, second, air traffic communications in China are conducted in the Mandarin dialect (Dennis, 2015a). This rationale was also used in Studies 4 and 5. All stimuli were recorded over a simulated propeller aircraft background noise, and the SNR was fixed. The SNR set up is described in section 6.2.4.2.

The equations used as stimuli consisted of two single- or double-digit integers with a stated sum based on the previous studies (e.g., Geary et al., 1993). The cognitive demands of the task were considered according to the complexity and difficulty of calculations rather than by differentiating between single- or double-digit integers. Hence, only single-operation mathematical problems that required calculation ability up to 20 (e.g., 11 + 6 = 17; 15 – 8

= 7) were presented. Over the three experimental conditions, each participant was therefore presented with 84 correct (distracting) stimuli and 46 incorrect stimuli, giving 130 stimuli in total. The higher proportion of distracting stimuli was designed to simulate real-life aviation communications, where infrequent errors are interspersed with more frequent, correct messages. All stimuli were different across the language conditions.

The 14 incorrect equations were randomly distributed amongst the 26 correct equations, giving 40 stimuli in total in each of the monolingual conditions (L1 and L2). In the Mix condition, the 18 incorrect equations were randomly distributed amongst the 32 correct equations, giving 50 stimuli in total. The probability of an incorrect equation being presented was approximately 35%, and the probability of a correct equation was approximately 65%. The number of additions and subtractions was equal; in each of the monolingual conditions (L1 and L2) there were 20 additions and 20 subtractions, and in the Mix condition, there were 25 of each mathematical operation. In the Mix condition, the stimuli were presented in both languages; 9 incorrect and 16 correct equations in Chinese, and 9 incorrect and 16 correct equations in English. The distribution of the stimuli was random, to minimize any learning or expectation effect.

In this study, participants were assessed on how quickly and accurately they responded to acoustic stimuli designed to represent correct and incorrect messages. According to Marslen-Wilson’s (1985) finding that words can be processed before they are fully heard, it was possible for participants to respond to stimulus while it was still playing. The RT (in seconds) was therefore measured from the onset of the auditory stimulus. However, due to the different lengths of Chinese and English spoken stimuli, the RT was then adjusted using the subtraction method explained in section 3.12, and pure RT was used for the subsequent data analyses. No visual stimuli were presented. The stimuli list is included in Appendix E.

6.2.4.2. Speech to Noise Ratio

The perception of speech stimuli, particularly in aviation, is also affected by external noise. In hindsight, the background noise level in Study 2 was low. This may have increased the intelligibility of the stimuli at that level, so that participants had a close to perfect hit rate with almost no false alarms or misses. Therefore, when developing Study 3 (and Studies 4 and 5), particular attention was paid to setting the background noise level by setting the SNR, which is the measure of audio signal level compared with the noise level present in the signal.

Only white noise was used in Study 2. White noise is defined as a random signal having equal intensity at different frequencies giving it a constant power spectral density (Green & Swets, 1966). Consequently, white noise has the capability to effectively mask other, especially distracting, sounds (Green & Swets, 1966). This characteristic is important in auditory psychophysics - white noise is often used for relaxation (Afshar et al., 2016). However, in aviation, noise produced by propellers during various phases of flight is a mechanical noise causing noise pollution. Because of the different noise characteristics of white noise and propeller noise and following the intention to simulate more aviation-like acoustic stimuli, propeller noise was used in this and subsequent studies. The propeller noise of a common general aviation airplane (Cessna) utilized frequently by flight crew and passengers. The main difference between pilots and non-aviation participants with regard the propeller background noise is the different duration of exposure to the sound, but not its familiarity. Nevertheless, the difference between the impact of the propeller background noise (used in this and the following studies) and the white noise (used in Study 2) on the results was explored using the discriminability index (*d'*). The discriminability index describes how discriminable was the signal (stimulus) from noise (Heeger, 2003). Thus, by comparing *d'* of the studies using two different background noises (the white noise and the propeller noise), the difference between the impact of these noises on performance can be investigated. The propeller aircraft in-flight interior noise was downloaded from the web page of Free Sound Effects FX Library (GRSites, n.d.) and added to the speech signal as a background noise.

Lamm and Lawrence (2010) investigated whether exposure to aircraft interior noise is a health hazard and found that the mean sound level during flights was 101.3dB. This is similar to the finding of Ericson and McKinley (2001) that the noise levels in most commercial aircraft cockpits range from 85 to 100dB. As pilots normally use a headset, which can mitigate the level of noise, it was also important to consider the noise levels with a headset. Lamm and Lawrence (2010) found that the use of a headset lowered the noise level by an average of 13dB. These considerations led to the set-up of the SNR. The equations were all spoken by a computerised female voice set to the same volume. All stimuli were normalized to the same decibel level to add the desired SNR. The SNR was defined as the ratio between the highest peak speech level plus 10 dB (speech) and the bottom of the noise range (noise). Thus, the speech signal (spoken equation) and noise were adjusted by the amplify effect setting using Audacity 2.1.0 as follows:

1. amplitude of the speech signal: 95dB
2. amplitude of the noise: 85 dB
3. new peak amplitude: 43.1 dB
4. gain of the noise: 0 dB
5. gain of the speech signal: +10 dB

Changing the gain of the speech signal track to +10 dB meant that the signal track played 10 dB louder than before. Applying the amplify effect setting to the values mentioned above caused a clipping distortion effect, which, in playback, created a "microphone effect" analogous to real aircraft radio communication with a static noise. Although amplification and gain did roughly the same thing (i.e., they influenced the volume level), there were some differences. Changing the gain did not affect the amplitude of the signal tracks, but changed the volume of the signal track, so the signal was more easily distinguished from the noise in terms of loudness. The amplify effect changed the amplitudes of the signal and noise in the tracks but did not affect the gain (volume) settings. Putting it together, amplifying the tracks of the signal and noise provided the difference in power, and setting up the gain of the tracks provided the difference in loudness (Steve, 2013).

The result after mixing the two tracks—spoken equation and noise of a propeller aircraft—was speech stimuli containing aircraft noise in the background.

6.2.4.3. English Language Proficiency

English language proficiency of the participants was evaluated using the IELTS Listening Test 1. The free IELTS Listening test was downloaded from the public website (British Council, n.d.d). The test took about 30 minutes to complete and participants were provided an additional 10 minutes to transfer their answers into the answer sheet, based on the instructions in the standard IELTS test.

6.2.5. Procedure

Participants were recruited either through an e-mail invitation through the Chinese Student Club at Massey University, personally on the Massey University campus, or through social media, between July 18 and October 31, 2016. After accepting the invitation, the participants were given a brief introduction to the study; they were informed that there were two tests—the IELTS Listening test to measure English language proficiency in listening and the language switching experiment itself—each lasting approximately 30 minutes, and that all instructions would be provided in English.

Participants started with the experiment to avoid any adverse effects on their performance caused by increased tiredness after performing the IELTS test. Because of the presence of the noise, participants first set up their own comfortable volume by following instructions for setting safe sound levels to protect their hearing. They were informed that if they felt any discomfort or found the volume levels too high at any time during the testing, they could adjust the volume, or switch off the sound source and stop the test immediately. Participants were given time to practice and fully comprehend the task. The task was to press '*yes*' on a keyboard when the equation was correct (e.g., 7 – 5 = 2), and '*no*' when it was incorrect (e.g., 5 + 5 = 12). The responses were reverse coded for the subsequent data analyses because the task objective was to identify errors; that is, in the analysis, the "yes" response referred to correct detections of erroneous equations (hits), and the "no" response referred to correct equations (correct rejections). This was identified as a limitation of the study.

Participants were able to press *Say again* as a measure of communication performance. Participants could choose the option *Say again* as many times as they needed, meaning that one stimulus could be replayed several times. During the experiment, there were no other acoustic stimuli, and no visual stimuli. The experiment lasted approximately 30 minutes, depending on the speed of responses and requests for stimulus repetition.

Following the experiment, participants performed the IELTS Listening Test 1, which took approximately 30 minutes. Based on the obtained scores they were assigned to one of four proficiency groups, in accordance with IELTS scoring practices (British Council, n.d.b): Intermittent–Modest user (*n* = 10; IELTS test scores: 2.5–5.5), Competent user (*n* = 14;

IELTS test scores: 6.0–6.5), Good user (*n* = 10; IELTS test scores: 7.0–7.5) and Very Good user (*n* = 6; IELTS test scores: 8.0–8.5).

Lastly, participants were asked to provide demographic information on their age, sex, mean duration of their stay in New Zealand, and whether they had any known hearing impairment. The study was conducted anonymously, so participants were not asked to provide their names or contact details, unless they wanted to receive a results summary. As gratitude for their effort and time, participants were provided with $5 worth of refreshments.

The software G*Power (Erdfelder, Faul, & Buchner, 1996) was used to determine sample size, considering the type of analysis to be performed, type and number of comparisons to be made, and the number of variables to be examined. Two power analyses were conducted for the two different designs. *A priori* power analysis for an omnibus 3-way ANOVA was conducted in which power was specified for the key comparison between the three language conditions. A total sample size of *n* = 28 was recommended for an experimental power of .80, with α = .05, and an effect size of *f* = .25 (based on Ison, 2011) in a repeated-measures, within-subjects design. *A priori* power analysis was also used to determine the total sample size in a repeated-measures between-subjects design with α = .05 and an effect size of *f* = .40 (based on Ison, 2011). There was an 81% chance of correctly rejecting the null hypothesis of no difference between the two English language proficiency groups (number of groups was based on the data from previous experiments) with a total of *n* = 30 participants.

6.3. Results

6.3.1. General findings

The data were screened for outliers. The screening identified only three outliers in the Language and ISI data, which were defined as z-scores greater than |3.29| as recommended by Tabachnick and Fidell (2007). After reviewing normal Q-Q plots following the same rationale as in Study 2 (see section 5.3.1), it was concluded that outliers might have a negative effect on findings, and they were excluded so that the data would more likely meet the requirements for the application of parametric statistical analysis. The assumptions required for the parametric statistical tests were tested, and where the assumptions were found to be violated, a non-parametric alternative was chosen. The level of statistical significance, alpha, was set at .05 for all statistical tests, and all tests were conducted as two-tailed.

Additionally, as the on-line method for measuring RT was used and a difference between the length of Chinese and English stimuli was observed, pure RT was calculated by subtracting the duration of each stimulus from its RT across all language conditions, and this was used for the analyses. The mean pure RT together with the mean durations of the stimuli are presented in Table 18.

Table 18

Mean Response Times (RT), Mean Durations of Stimuli, and Mean Pure RT in Seconds, in First (L1), Second (L2), and Language Switching (Mix) Conditions

	Language condition					
	L1	*SD**	L2	*SD*	MIX	*SD*
Mean RT	3.382	0.724	3.763	0.738	3.637	0.588
Mean duration of stimuli	2.029	0.253	2.200	0.203	2.168	0.270
Mean pure RT	1.353	0.723	1.563	0.738	1.469	0.588

*SD = standard deviation

To provide a comprehensive analysis, and for clarity, the Results section is organised under two main indices of performance; speed and accuracy. Each section is further organised according to the within- and between-subjects variables. SDT measures are presented at the end of the Results section.

6.3.2. Performance Speed

Multivariate analysis was conducted for the primary interest of the comparison of performance speed across the different language conditions. A within-subjects 3 × 3 ANOVA was used to explore the effect of Language condition and ISI factors on speed of performance. The Language condition factor consisted of three levels (L1, L2 and Mix), as did the ISI factor (1 s, 4 s, and 9 s). As the repeated-measures ANOVA is not robust to violations of sphericity (Schmider et al., 2010), Mauchly's test of sphericity was considered first. The assumption of sphericity had been violated for the main effects for Language condition ($\chi^2(2) = 28.283$, $p < .001$) and ISI factors ($\chi^2(2) = 6.258$, $p = .044$), and their interaction ($\chi^2(9) = 17.221$, $p = .046$). Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity for the Language condition factor (ε = 0.652), and Huynh-Feldt estimates of sphericity for the ISI factor (ε = 0.903) and for the interaction effect (ε = 0.863). This was decided based on the values of Epsilon (ε = .75), according to Girden (1992).

There was evidence of statistically significant main effects for the Language condition (Greenhouse-Geisser adjusted $F(1.303, 49.531) = 4.890$, $p = .023$, $\eta_p^2 = .114$), and for the ISI factors (Huynh-Feldt adjusted $F(1.806, 68.623) = 8.019$, $p = .001$, $\eta_p^2 = .174$). The interaction effect was also statistically significant (Huynh-Feldt adjusted $F(3.453, 131.211) = 2.984$, $p = .027$, $\eta_p^2 = .073$). The interaction plot is presented in Figure 7.

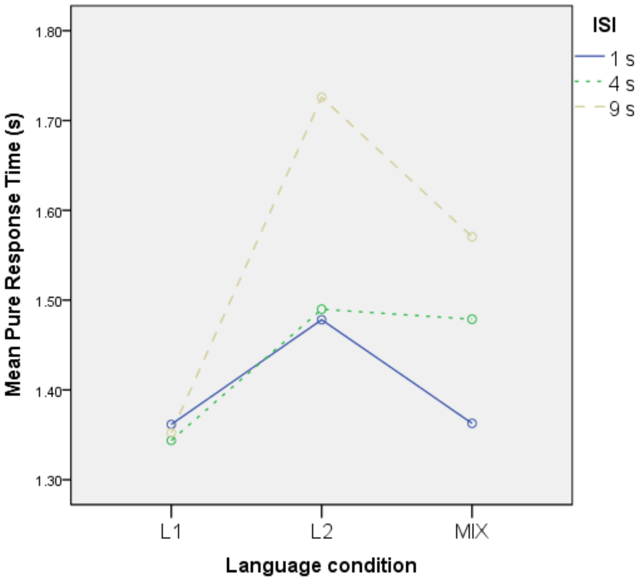


Figure 7. Interaction plot for Language condition and ISI factors.

6.3.2.1. Performance Speed and Language Conditions

Prior to *post hoc* analysis of the effect of the Language condition factor on performance speed, the assumption of normality was tested and found to be satisfied, as the skewness and kurtosis levels of some of the data were between the span of –2.0 and +2.0 (Cramer, 1998; George & Mallery, 2010). A paired samples *t*-test was conducted to compare the mean pure RTs across the language conditions (L1, L2 and Mix), which indicated statistically significantly faster performance in the L1 condition (*M* = 1.28, *SD* = 0.54) than in the L2 condition (*M* = 1.56, *SD* = 0.75) (*t*(38) = –2.371, *p* = .023, *d* = .22), or the Mix condition (*M* = 1.45, *SD* = 0.58) condition (*t*(38) = –2.129, *p* = .040, *d* = .20). There was no evidence of a statistically significant difference in performance speed between the L2 and the Mix condition (*p* = .184).

6.3.2.2. Performance Speed and ISI

The test of the assumption of normality of the differences between the interactions of levels of Language condition and ISI factors was considered to be violated as the skewness and kurtosis levels were beyond the span of –2.0 and +2.0 (Cramer, 1998; George & Mallery, 2010). A non-parametric Wilcoxon signed-rank test was therefore conducted, which suggested that there was no performance difference in the L1 condition explained by ISI (*p* ≥ .748). In the L2 condition, the performance on ISI 9 s was slower than on ISI 4 s (*Mdn* = 1.67 vs. *Mdn* = 1.35; *Z* = –2.997, *p* = .003, *r* = .11), and also slower than on ISI 1 s (*Mdn* = 1.67 vs. *Mdn* = 1.39; *Z* = –2.634, *p* = .008, *r* = .10). In the Mix condition, the only difference was between ISI 9 s and ISI 1 s (*Mdn* = 1.59 vs. *Mdn* = 1.32; *Z* = –2.42, *p* = .016, *r* = .09). The performance difference between the L1 and the L2 conditions was found only when the stimuli were separated by the longest ISI 9 s (*Z* = –3.04, *p* = .002, *r* = .11). The performance also differed between the L1 and the Mix language conditions in both ISI 4 s (*Z* = –2.19, *p* = .028, *r* = .08), and ISI 9 s (*Z* = –2.07, *p* = .039, *r* = .08). Other interactions were not statistically significant at .05. Table 19 summarises the data for the interaction effect.

Table 19

Median (Mdn) Pure Response Times in Seconds across Language Condition and Inter-stimuli Interval (ISI) Factors

	ISI 1 s	ISI 4 s	ISI 9 s
	Mdn	*Mdn*	*Mdn*
L1	1.079	1.070	1.114
L2	1.390	1.352	1.666
Mix	1.316	1.361	1.586

6.3.2.3. Performance Speed on Correct and Incorrect Equations

To test the second research question of how quickly participants could identify erroneous equations compared with correct equations, an assumption of normality of distribution was tested and was considered to be violated, as the skewness and kurtosis levels were beyond the span of –2.0 and +2.0 (Cramer, 1998; George & Mallery, 2010). Therefore, a non-parametric Wilcoxon signed-rank test was used, which indicated that regardless of the language condition, it took longer to correctly identify erroneous equations (to make *hit* type of responses) than to correctly respond to correct equations (to make correct rejections, *CR*); in the L1 (*Z* = –4.95, *p* < .001, *r* = .32), L2 (*Z* = –3.43, *p* = .001, *r* = .22), and Mix conditions (*Z* = –2.81, *p* = .005, *r* = .18). There was no evidence of a statistically significant difference between the L2 and Mix conditions when making CR (*p* = .757), or hits (*p* = .50). Neither was there a statistically significant difference between the L2 and the L1 (CR: *p* = .088, hits: *p* = .158). The only difference was found between the L1 and the Mix condition when correctly responding to correct stimuli (*Z* = -2.81, *p* = .005, *r* = .18), indicating slower responses in the Mix condition (*Mdn* = 1.43) than in the L1 condition (*Mdn* = 1.01). Median pure RTs are reported in Table 20.

Table 20

Median (Mdn) Pure Response Times in Seconds on Correct Rejections (CR) and Hits across the Language Condition Factor

	CR	Hits
	Mdn	*Mdn*
L1	1.01	1.43
L2	1.46	1.58
Mix	1.43	1.35

6.3.2.4. Switch Costs and Mixing Costs

Given the violated assumption of normality of the differences in switch costs and mixing costs data, based on the skewness and kurtosis levels that were beyond the span of –2.0 and +2.0 (Cramer, 1998; George & Mallery, 2010), a non-parametric test was used.

There was a statistically significant difference between RTs to Chinese and English stimuli in the Mix condition (mixing costs), based on Friedman’s test ($\chi^2(7) = 44.533$, $p < .001$). A Wilcoxon signed-rank test indicated that responses on English stimuli after hearing a stimulus in Chinese (*Mdn* = 1.46) took longer than responses on Chinese stimuli after hearing a stimulus in English (*Mdn* = 1.23) when the equations were correct ($Z = –2.890$, $p = .004$, $r = .16$). No mixing costs were found when the equations were incorrect ($p = .778$). In other words, there was no difference in performance speed in the Mix condition on either Chinese or English stimuli when identifying errors. A comparison of RTs between CR and hits when switching to Chinese stimuli in the Mix condition revealed a statistically significant difference ($Z = –3.387$, $p = .001$, $r = .19$). When switching to English stimuli, there was no difference between CR and hits ($p = .657$). Table 21 summarises the data for the switch costs and mixing costs.

Table 21

Median (Mdn) Pure Response Times in Seconds on Chinese and English Word Stimuli for Correct Rejections (CR), Hits and Both Combined in Monolingual (L1 and L2) and Language Switching (Mix) Conditions; Mixing Costs and Switch Costs

	Monolingual		Mix			Switch costs
	Sequence	*Mdn*	Sequence	*Mdn*	Mixing costs	(Pure vs. Mix)
CR	L1 → L1	1.01	L2 → L1	1.23	.23	→ L1: .22
	L2 → L2	1.37	L1 → L2	1.46		→ L2: .09
Hits	L1 → L1	1.39	L2 → L1	1.38	.19	→ L1: .01
	L2 → L2	1.58	L1 → L2	1.57		→ L2: .01
CR and hits	L1 → L1	1.12	L2 → L1	1.23	.07	→ L1: .11
Combined	L2 → L2	1.50	L1 → L2	1.30		→ L2: .20

A Wilcoxon signed-rank test indicated statistically significant switch costs between responses on English stimuli in the L2 condition and responses on English stimuli in the Mix condition (Z = –2.258, p = .024, r = .13). There was evidence of longer RTs on English stimuli in the Mix condition (*Mdn* = 1.46) than in the L2 condition (*Mdn* = 1.37), when the equations were correct. However, no statistically significant difference was found between RTs on stimuli in either of the monolingual conditions (L1 and L2) in comparison to the corresponding language stimuli in the Mix condition, when correctly identifying an error ($p \geq .354$).

Finally, when CR and hits were not distinguished in the analysis but were combined as an overall RT, a non-parametric Friedman's test of differences indicated that there was no significant difference between switch or mixing costs ($\chi^2(3) = 6.99$, p = .72).

6.3.2.5. Effect of English Language Proficiency on Performance Speed

The descriptive statistics associated with participants' mean pure RTs across the four English language proficiency groups are reported in Table 22. Both the assumption of normality for a between-subjects ANOVA and the assumption of homogeneity of variances were evaluated and were found to be satisfied as the four group distributions were associated with skewness and kurtosis within the span of –2.0 and +2.0 (Cramer, 1998; George & Mallery, 2010), and based on Levene's *F* test, verified in the L2 condition (*F*(3, 36) = 1.805, *p* = .164), and in the Mix condition (*F*(3, 36) = 0.339, *p* = .797). Therefore, a between-subjects ANOVA was performed, which indicated statistically significant differences between the English language proficiency levels in both the L2 (*F*(3, 36) = 3.087, *p* = .039, η_p^2 = .205), and Mix conditions (*F*(3, 36) = 3.182, *p* = .035, η_p^2 = .210).

Table 22

Mean Pure Response Times in Seconds across the English Language Proficiency Levels

	IELTS Listening	*n*	L2	*SD*	Mix	*SD*
Intermittent–Modest user	2.5–5.5	10	1.86	0.91	1.68	0.60
Competent user	6.0–6.5	14	1.74	0.62	1.65	0.56
Good user	7.0–7.5	10	1.44	0.66	1.33	0.51
Very Good user	8.0–8.5	6	.87	0.31	.93	0.43

The highest level of English language proficiency (Very Good user) was associated with the quickest mean pure RT (L2: *M* = .87; Mix: *M* = .93), and the lowest level of English language proficiency (Intermittent–Modest user) was associated with the slowest mean pure RT (L2: *M* = 1.86; Mix: *M* = 1.68). Pairwise comparisons of the means using the Least Significant Difference procedure indicated that statistically significant differences lay between the Intermittent–Modest user (*M* = 1.676) and Very Good user (*M* = 0.931; *p* = .012), and between the Competent user (*M* = 1.652) and Very Good user (*M* = 0.931; *p* = .010) in the Mix condition, and between the Intermittent–Modest user (*M* = 1.857) and Very Good user (*M* = 0.866; *p* = .008), and between the Competent user (*M* = 1.736) and Very Good user (*M* = 0.866; *p* = .013) in the L2 condition. There was no evidence of a significant difference between other English proficiency levels in the examined task at the .05 significance level. Data are presented in Figure 8.

Note that three outliers are displayed in the L2 condition in Figure 8. The reason is that SPSS, in Boxplot analysis, defines two different types of outliers, based on two different inter-quartile range (IQR) rule multipliers: 1.5 IQR's values are denoted with a circle, and 3 IQR's values are denoted with an asterisk (IBM Knowledge Center, n.d.). However, according to Hoaglin and Iglewicz (1987), the 1.5 multiplier is inaccurate. Therefore, as suggested in section 6.3.1, outliers in this thesis were defined according to the Tabachnick and Fidell (2007) recommendation, as z-scores greater than |3.29|, and confirmed using Q-Q plots.

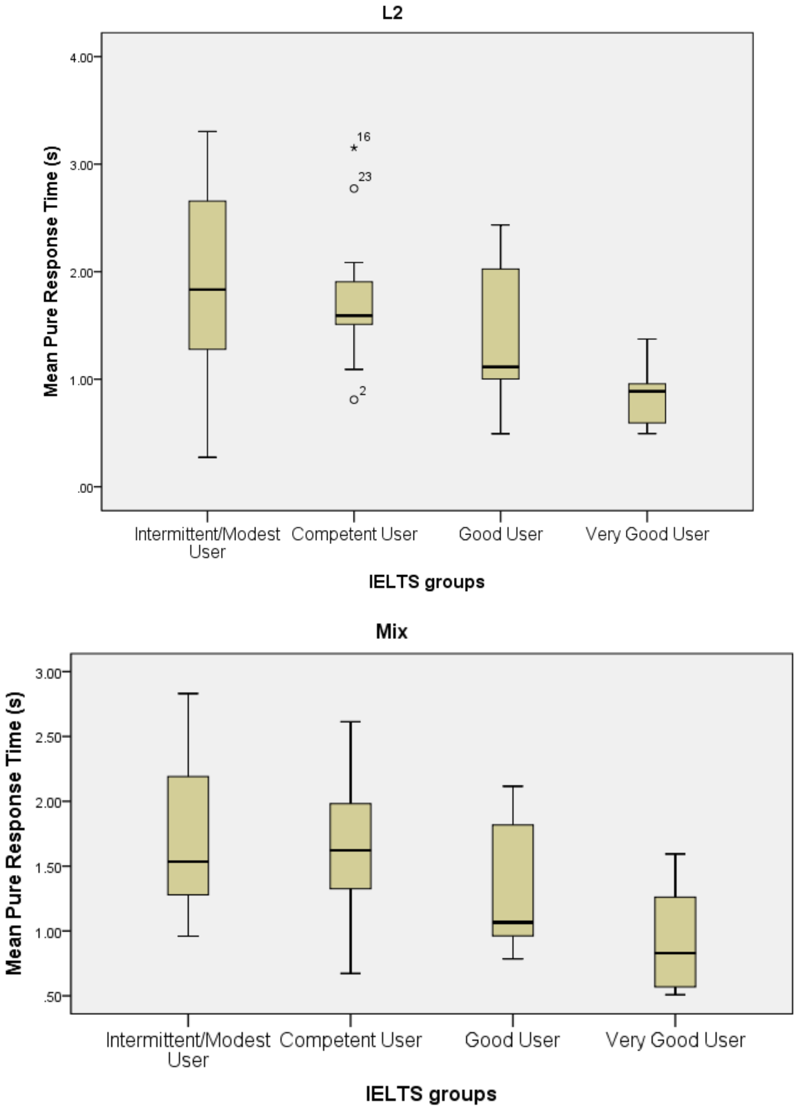


Figure 8. Profile plots for the English language proficiency groups in the L2 and Mix conditions.

6.3.3. Performance Accuracy

6.3.3.1. Performance Accuracy, Number of Repetitions and Language Conditions

Prior to testing whether performance accuracy differed between language conditions, the assumption of normality was evaluated and was found to be violated, as the distributions were associated with skewness and kurtosis beyond the span of –2.0 and +2.0 (Cramer, 1998; George & Mallery, 2010). The differences in performance accuracy between language conditions were found to be statistically significant, based on Friedman's test ($\chi^2(2) = 27.569$, $p < .001$). Data are presented in Table 23.

A Wilcoxon signed-rank test further indicated greater false alarm rates (responding to correct equations as if they were incorrect) than miss rates (missing incorrect equations) in the L2 condition (9.81% vs. 3.75%; $Z = –4.191$, $p < .001$, $r = .66$, risk ratio = 2.62) and the Mix condition (7.58% vs. 3.47%; $Z = –3.967$, $p < .001$, $r = .63$, risk ratio = 2.18). The false alarm rates were found to be significantly smaller in the L1 condition than the L2 condition (2.12% vs. 9.81%; $Z = –4.375$, $p < .001$, $r = .69$, risk ratio = 4.64), and the Mix condition (2.12% vs. 7.58%; $Z = –4.800$, $p < .001$, $r = .76$, risk ratio = 3.58). No other comparisons were statistically significant.

Table 23

Error Types (Miss and False Alarm), Hits and Correct Rejections (CR), Total Number of Errors and "Say Again" Requests across Language Conditions, and Percentage of Errors from 5200 stimuli (%Error$_T$)

	Miss	False alarm	Hits	CR	Error Total	%Error$_T$	Say again
L1	12	22	548	1018	34		149
L2	21	102	539	938	123		604
Mix	25	97	695	1183	122		569
Total	58	221	1782	3139	279	5.37	1322

Data for the number of *Say again* requests are also summarised in Table 23. Friedman's test indicated that the number of *Say again* requests also differed significantly across the language conditions ($\chi^2(2) = 19.234$, $p < .001$). A Wilcoxon signed-rank test further indicated that the number of requests to replay a stimulus was four times higher in the L2 condition than the L1 condition ($Z = –2.757$, $p = .006$, $r = .25$, risk ratio = 4.054) and almost four times higher in the Mix than in the L1 condition ($Z = –4.309$, $p < .001$, $r = .39$, risk ratio = 3.819). There was no evidence of a statistically significant difference between the L2 and the Mix conditions ($p = .808$).

6.3.3.2. Performance Accuracy and ISI

Prior to testing whether performance accuracy differed between the ISIs, the assumption of normality was evaluated and was found to be violated as the distributions were associated with skewness and kurtosis beyond the span of –2.0 and +2.0 (Cramer, 1998; George & Mallery, 2010). Friedman's test was conducted and indicated statistically significant differences ($\chi^2(2) = 7.554$, $p = .023$). Paired comparisons using Wilcoxon signed-rank tests indicated that the error rates were greater on ISI 4 s than on ISI 1 s (5.99% vs. 3.47%; $Z = –2.052$, $p = .040$, $r = .32$, risk ratio = 1.73), and greater on ISI 9 s than ISI 1 s (6.79% vs. 3.47%; $Z = –2.653$, $p = .008$, $r = .42$, risk ratio = 1.96). There was no difference in performance accuracy between ISI 4 s and 9 s ($p = .128$). Data are summarised in Table 24.

Table 24

Number (n_{errors}) and Percentage of Errors (%Error$_E$) across ISI Levels and as a Proportion of Total Number of Stimuli ($n_{stimuli}$) in Each ISI Level (%Error$_{ISI}$)

	ISI 1 s	ISI 4 s	ISI 9 s
n_{errors}	62	103	114
%Error$_E$	22.22	36.92	40.86
%Error$_{ISI}$	3.44	5.99	6.79
n_{stimuli}	1800	1720	1680

6.3.3.3. Effect of English Language Proficiency on Performance Accuracy

Prior to testing whether performance accuracy differed between the English language proficiency levels, the assumption of normality was evaluated and found to be violated as the distributions were associated with skewness and kurtosis beyond the span of –2.0 and +2.0 (Cramer, 1998; George & Mallery, 2010). Therefore, a non-parametric Kruskal-Wallis H test was performed, which indicated statistically significant differences in performance accuracy across the three language proficiency groups ($H(3) = 11.101$, $p = .011$).

Further analyses using Mann-Whitney tests indicated that Very Good users made significantly fewer errors than Competent users ($U = 10.00$, $p = .006$, $r = .30$, risk ratio = 2.201), but there was no evidence of a significant difference between Very Good and Good ($p = .696$) or Intermittent–Modest users ($p = .081$). Good users made fewer errors than Competent users ($U = 24.00$, $p = .007$, $r = .30$, risk ratio = 1.441). There was no evidence of a significant difference between Competent and Intermittent–Modest users ($p = .906$). Data are summarised in Table 25.

Table 25

Number of Participants (n), Correct Responses and Errors, and Percentage of Errors (%Errors) across English Language Proficiency Levels

	IELTS Listening	*n*	Correct	Errors	%Errors
Intermittent–Modest user	2.5–5.5	10	1202	98	7.538
Competent user	6.0–6.5	14	1707	113	6.209
Good user	7.0–7.5	10	1244	56	4.308
Very Good user	8.0–8.5	6	758	22	2.821

6.3.4. SDT Measures

To determine whether the responses stemmed from different internal criterions used for decisions (*C*) or from the increased level of noise, sensitivity (*d'*) and response bias were calculated. The findings revealed that despite the perceived disturbing effect of noise reported by participants, the discriminability of the signal was very high in all language conditions (*d'* ≥ 3.07). There was a small bias towards *yes* responses, and thus a tendency to false alarms, in both the Mix and the L2 conditions (*C* ≥ –0.19). In the L1 condition, the responses were almost unbiased (see Table 26).

Table 26

Sensitivity (d') and Decision Criterion (C) of Error Identification Task across the Three Language Conditions: Native (Chinese) Language (L1), Second (English) Language (L2), and Language Switching (Mix)

	z(HR)	*z(FAR)*	*d'*	*C*
L1	2.03	–2.03	4.06	–0.000
L2	1.78	–1.29	3.07	–0.245
Mix	1.82	–1.44	3.26	–0.190

6.4. Discussion

To answer the first guiding research question, performances between the language conditions were compared. The finding of fastest responses in the L1 condition is somewhat intuitive. The task required comprehension of equations, which might be dependent on language dominance and language proficiency (e.g., Pavlenko, 2014; Tamamaki, 1993). Yet, neither error identification performance speed nor accuracy differed between the L2 and the Mix condition. Despite this, the switch cost analysis revealed longer RTs on English language stimuli in the Mix condition than in the L2 condition when participants were responding to correct equations. This can suggest that the identification of erroneous equations likely minimized the difference in language control demands, and the difficulty of the process of identifying an error was the primary factor affecting performance speed.

There was evidence of responses on second language stimuli after hearing a stimulus in the native language in the Mix condition being slower than responses on native language stimuli after hearing a stimulus in the second language. This indicates the presence of symmetric mixing costs, which is contrary to the findings of previous research (Bobb & Wodniecka, 2013; Costa & Santesteban, 2004a; Meuter & Allport, 1999). However, this effect was observed only when responding to correct equations. When responding to incorrect equations, there was no difference in processing speed on native and second language stimuli in the Mix condition. This may support the explanation provided above, that the influence of cognitive demands on error identification is greater than the language processing demands.

The difference between the cognitive demands of error identification and of language switching might be explained according to Smith and Stein’s (1998) levels of cognitive demands (which were developed originally for dealing with mathematical tasks). That is, the lowest-level cognitive demand tasks require simple memorization, whereas higher-level cognitive demand tasks require mathematical problem solving. In this study, to identify an erroneous equation, participants had to use mental computation, arguably a higher-level demand. Whereas, when perceiving different languages, the language processing of spoken word recognition is given by the input words; that is, phonological identification of a target language (Costa & Santesteban, 2004b). This suggests somewhat passive process in a sense that a listener does not actively choose the target language. However, it remains an open question as to why error identification caused longer latencies than responses to correct equations, provided that mental calculation shall be used in both cases.The reliability of the measurement of pure RT also needs to be discussed (its discussion in this section parallels the discussion in Study 2). A question that could verify the reliability of the suggested approach to RT measurement in language switching studies was proposed in Study 2; whether it is possible to observe longer latencies in the L2 condition, when the length of the second language stimuli is generally longer than that of the native language stimuli (see section 5.4). The findings of this study suggested that, in fact, it is possible; performance was found to be fastest in the L1 condition. However, this still cannot be considered proof, because an empirical comparative study would be required to provide sufficient explanation.

Nonetheless, although performance was found to be fastest in the L1 condition, there was no difference in speed of performance between different ISIs. However, performance speed was affected by ISI factors in both the L2 and Mix conditions; specifically, the longest interval (9 s) between the stimuli invoked the slowest RT. The findings also indicated that the number of errors increased after the shortest (1 s) interval, but there was no difference in performance accuracy between the medium (4 s) and long (9 s) interval. This is somewhat contrary to the previous finding that the lowest probability of signal detectability was immediately after the signal observation (Daniel & Pikala, 1976). In other words, more errors were observed in the shortest ISI. This was attributed to the least expectation of another signal. Unfortunately, there is insufficient information to discuss the observed discrepancies in further detail, and more research would be needed to make further assumptions regarding the ISI.

The performance was also found to be the most accurate in the L1 condition. No evidence of a difference in accuracy between the L2 and the Mix condition was found. Although greater false alarm rates (responding to a correct equation as if it was incorrect) than miss rates (missing an erroneous equation) suggesting a bias towards *yes* responses in the Mix and the L2 conditions, it took longer to correctly identify the presence of an error than correctly respond to the correct stimuli. This can be attributed to the nature of the task, which was to identify an error. Presumably, when participants were instructed to identify an error, it could affect their expectation of errors and, thus, their bias. Interestingly, however, this bias was observed only in the L2 and the Mix conditions. In the L1 condition, the responses were almost unbiased. Thus, the second research question was answered. A potential implication for the SA of aviation personnel is that error identification performance can be positively affected by a reasonable expectation of errors.

Another question was whether the dominance of the native language affected the confidence with which participants responded, and thus affected their response bias. Analysis of the number of requests to *Say again* between the conditions can provide some support for the effect of language dominance. A request for stimulus repetition may suggest that a participant was either uncertain of what was perceived, or uncertain about the correctness of the equation. Using this rationale, the lowest occurrence of *Say again* found in the L1 condition could therefore indicate the highest confidence in decisions and responses; on average, four times greater than in the L2 or in the Mix condition.

Additionally, 1.12% (n = 58) of all errors were miss type of errors. This can be reviewed within the real-life findings of Prinzo et al.’s (2006) study, in which 8% of all read-back errors were corrected. The Introduction section included a discussion of whether there is a difference between errors of non-correction and the errors of non-detection. Prinzo et al. (2006) suggested that the errors were detected but were not corrected because of a conservative process, with corrections reserved for transmissions that had a direct or immediate effect on safety. The present findings, however, suggest that some erroneous transmissions can be missed by an operator. It is reasonable to expect even greater occurrence of such errors in real life, given that this rate (1.12%) occurred despite there being an expectation bias—a tendency towards *yes* responses—likely invoked by the instruction. However, in real life, where attention is not directed towards potential errors, an opposite bias might be employed; a confirmation bias. The confirmation bias is a "tendency to look for information to confirm a decision already made" (FAA, 2008, p. 3). For example, Monan (1988) found that pilots had an implicit expectation that a lack of an ATCO's response was a silent confirmation that the read-back was correct, which, in turn, reduced pilots' own active listening. This can have critical implications for safety. The assumption of non-correction can just reinforce this implicit expectation (or confirm read-back accuracy), and thus increase the possibility of adverse consequences.

However, performance accuracy also needs to be reviewed with consideration given to the background noise. The acoustic stimuli had relatively high noise levels, 95 dB. According to Lehto and Landry (2013), loud noise can cause annoyance and speech interference. With a noise level of 85 dB, the average annoyance rating was about 4.2 on a five-point Likert scale (Lehto & Landry, 2013), indicating that participants found this level of noise annoying. From their regression analysis, Lehto and Landry (2013) developed an equation that describes the percentage of words missed at different noise levels, with about 5.7% of the words being missed because of noise interference at 95 dB. The percentage of errors was almost equal to that of the current study (5.37%; see Table 23). A question, therefore, arose as to whether findings related to accuracy can be explained by the level of noise—95 dB, which could be considered annoying, based on Lehto and Landry's (2013) findings—rather than as an effect of the language condition. A potential answer can be found in the SDT measures—the discriminability index, *d'*. The discriminability (*d'*) of the speech was found to be greater than 3.0, indicating that despite the perceived disturbing effect of noise, the discriminability of a signal was still very high in all language conditions.

There was some evidence of an effect of English language proficiency. The size of the effect was very similar between the L2 and Mix conditions. An increased level of English language proficiency caused the number of errors to decrease. However, this effect was observed only at the Competent level of the IELTS Listening score (6.0–6.5). No difference in performance accuracy was observed between the Competent users and Intermittent–Modest users, nor between the Good and Very Good users. Therefore, a cut-off score of 6.0–6.5 (or Competent) on the IELTS Listening test could be considered for better performance in terms of more accurate responses. This is consistent with the attempts to align the IELTS and ICAO rating scales (see section 5.2.4.2.1), where the minimum requirement for Operational level 4 was estimated to be comparable to IELTS level 6.0 (based on Cambridge English, 2016; Harcourt Assessment, Inc., 2006).

There are three potential limitations of the current study. First, the task did not precisely simulate the real-life errors in air traffic communication messages as this study presented a simplified analogy of problem solving involving numbers. Whether similar results would be obtained in the identification of erroneous read-backs cannot be determined. Therefore, caution must be exercised when generalizing the findings. Second, a different type of background noise was used in this study, compared to Study 2. Different types of background noise add the potential for difficulties in comparing between studies that used white noise and studies that used propeller noise. However, this limitation can be addressed in the following studies by the use of the same propeller background noise. Third, caution also needs to be applied when generalizing the findings to the aviation population because the participants were not aviation students. However, as discussed in the previous Study 2, this might not be particularly problematic, given the nature of language processing being similar among different populations with inter-individual variability caused by the degree of the effect (Orlady & Orlady, 1999).

It would be interesting in future research to analyse error identification when participants do not expect an error to occur, and thus test the identification of infrequent errors under more realistic conditions. The analysis could be extended to alerts issued by automatic systems, beyond pilot–controller communications. Testing pilots and ATCOs’ detection of errors in automated systems could facilitate exploration of the human–machine interface.

To conclude, the findings of this study suggested that error identification performance was fastest and most accurate in the L1 condition. No difference was found between the L2 and Mix conditions. This, however, would be the most relevant finding for the consideration of bilinguals’ SA in bilingual versus monolingual air traffic environments. The finding that participants took longer to identify a mistake (though with greater accuracy) than to respond to a correct equation (albeit with more mistakes), provide general implications for response bias in aviation safety, rather than specific implications for SA.

CHAPTER SEVEN

Study 4: Prediction

7.1. Introduction

While most of the research on SA has centred on the first level (recognition), possibly because its investigation is the easiest (Klein, 2000), only a small number of studies have been conducted on prediction. Most likely this is because there is no known metric for knowing if an operator is correctly predicting future events.

The third level of SA has been studied less than the other two levels (Boudes & Cellier 2000; Sulistyawati, Wickens, & Chui, 2011). There may be several reasons, which could be summed up into three inter-related limitations (Boudes & Cellier, 2000): (i) *anticipation* is not an end in itself—the outcome behaviour or act, but rather an inner process; (ii) it generally does not result in directly observable behaviour; and, most importantly, (iii) *prediction* is not always indicated by explicit verbal expressions. According to Hoc (as cited in Boudes & Cellier, 2000), less than 4% of verbal expressions had to do with the future. However, research in this thesis is based only on perceptions of verbal messages, to develop SA in different language contexts. These three limitations put constraints on measuring the outcome behaviour using objective indices of performance, such as RT and errors.

As can be seen above, Boudes and Cellier (2000) used two terms: 'anticipation' and 'prediction'. In fact, the first challenge that needs to be addressed relates to the ambiguity within the terminology used to describe the third level of SA. This, along with the three limitations proposed above, will be discussed in the following sections. Attention will be focused on aspects that apply to the study of prediction within the context of bilingual air traffic environment.

7.1.1. Many Terms, Same Concept? Anticipation, Prediction, Expectation, Projection

Four different terms (anticipation, prediction, expectation and projection) are often used interchangeably when investigating the third level of SA, but they are rarely terminologically differentiated. For example, Banbury et al. (2004) defined level 3 SA as the ability to anticipate, think ahead or project. In Endsley's definition of the level 3 SA (2000; cf. Jackson, Chapman, & Crundall, 2009), however, these terms are used not as synonyms, but as an ability or a goal. Endsley (2000, p. 7) wrote: "This ability to project from current events and dynamics to anticipate future events (and their implications) allows for timely decision making". This suggests that the third level can consist of two parts—projection and anticipation. This understanding would be in accordance with the CAA's (2014, p. 71) definition as "being able to project ahead to predict what is likely to happen next". According to Bubic, von Cramon, and Schubotz (2010), although these and other similar terms are used with respect to predictive processing, they do not necessarily convey the same meaning.

Definitions of the terms typically used with respect to predictive processing are, therefore, first reviewed. Unfortunately, even in the *Dictionary of Psychology* (Chaplin, 1985), one term was defined by the other. For example, *projection* was defined as "a prediction beyond a given data" (Chaplin, 1985, p. 358), and *prediction* as "a statement about an event with respect to its future outcome" (Chaplin, 1985, p. 349). However, *anticipation* was not defined as a process but rather as "a mental set, or readiness, to receive a stimulus" (Chaplin, 1985, p. 30), and *expectation* was defined as "a state of anticipation" or "an emotional attitude of watchful waiting" (Chaplin, 1985, p. 166). It appears, then, that only prediction and projection refer to predictive processes. Additionally, projection is similarly defined as *extrapolation*, which stands for estimating a variable beyond the given data, and usually "takes the form of extending a curve beyond its plotted range" (Chaplin, 1985, p. 169), such as the future trajectory of an aircraft.

A terminologically ambiguous situation is potentially made even more challenging by prediction being understood differently in language switching and aviation SA studies. In language switching studies, the term anticipation is used, and it is analysed within speech comprehension (e.g., Foucart et al., 2014). In aviation, particularly in SA measurement, comprehension and anticipation are distinguished as two different levels. Next, the tasks used for the analysis of anticipation in linguistic studies consist of sentence, or missing word, completion (e.g., Foucart et al., 2014). In contrast, in aviation, anticipation was analysed as the prediction of an aircraft future position (e.g., Carretta, Perry, & Ree, 2009; Palacios, Doshi, & Gupta, 2008). A promising solution to these discrepancies was found in an operator-focused cognitive approach to SA, outlined in section 3.2. Of relevance is the laboratory-based paradigm used in aviation, which requires participants to predict the next stimulus position of a repeating sequence (Banbury et al., 2004).

Banbury et al. (2004) used a concept of *transition* within a cognitive streaming account of SA. Transitional probabilities refer to the likelihood that an event will occur following the occurrence of other events. Banbury et al. (2004) provided the following explanatory example. In a sequence of random numbers, each digit gives no inherent information about the next in the sequence and therefore, the transitional probabilities are low, because there are no linkages between them. On the other hand, when transitional information is present in a sequence as a familiar pattern or rule (e.g., 3, 6, 9, etc.), the transitional probability for the number "12" is high. This is the classical example of the number series test. In this case, an individual is extending the numerical sequence of numbers, and thus uses the process of extrapolation.

Transitions can explain the temporal nature of SA and why SA is usually acquired over time rather than instantaneously, based on the observed dynamics of a situation. Applying the transitional information forward allows prediction of a future state. In aviation, an ATCO can track the movement of an aircraft and predict its likely future position based on its trajectory (Banbury et al., 2004). By observing its movement over time, the ATCO can identify the type of pattern or transitional information, thus gaining an indication of how it is likely to behave in the future. The process of acquiring SA begins when the aircraft first appears on the radar screen (Banbury et al., 2004), or analogously, when the first digit of a number sequence is heard.

A similar approach was used in previous research (for a review, see Croft et al., 2004). In a laboratory-based paradigm, participants were required to predict the next stimulus position of a repeating sequence. Level 3 SA was then indicated by the ability of pilots to correctly predict a manner of future attack of an aircraft based on their experience in a previous simulator exercise. Lomov et al. (1983) found the time of decision making to be dependent on the logical conditions that an operator had to verify. It was found that with an increasing number of logical conditions the probability of making a mistake increased. When considering 3–4 logical conditions, fast and accurate actions become very complicated (Lomov et al., 1983).

To sum up, although this terminological differentiation is rather theoretical, it was included to develop the experiment, for which a definition of the measured phenomenon must be clarified, especially if the third level of SA is to be measured by the corresponding, underlying cognitive process. Without these insights on the terminological consensus, it would not be possible to develop proper measurement of level 3 SA.

7.1.2. Number Series

Of the tests of cognitive abilities, according to Sulistyawati, Wickens and Chui (2011), a math aptitude test was highly associated with prediction. Kunde (2005) suggested that the focus on quantitative prediction means that future events are predicted given a sequence of observations over time. This can be expressed using a number series test. As was suggested in the previous section, a number series test can reflect the ability to extrapolate the transitional information of the observed events to identify the future state.

Number series present numerical sequences that follow a logical rule, or pattern, which is based on elementary arithmetic. An initial sequence is given from which the pattern is to be deduced and, then, the participant predicts the next number that obeys the pattern. Alternatively, participants can be first provided with a certain number, the occurrence of which they need to predict in a given sequence of numbers that follows. The pattern can represent the dynamic changes in an environment, which are an important feature of SA. The goal is to predict what comes next, before it actually happens. For example, Allendoerfer et al. (2008) found that ATCOs continually search for potential hazards, and take actions proactively; that is, they respond before an automation system activates an alert.

Based on the above review, a number series test was chosen as the task to measure the underlying cognitive process of level 3 SA. Number series tests are also a common aptitude test used in the psychometric assessments of professional pilot candidates, such as the full ADAPT assessment developed by Symbiotics Ltd. (n.d.). The ADAPT is an online test consisting of five stages that provides information, among others, on the candidate's cognitive skills (Symbiotics Ltd., n.d.). Therefore, this study can be beneficial both theoretically, in researching the level 3 SA in a bilingual versus monolingual contexts, and practically, as a measure of the cognitive skills required for successful professional pilot candidates.

The two guiding research questions for Study 4 were:

Question 4a: Do speed and accuracy of prediction differ between monolingual and bilingual conditions?

Question 4b: How many steps ahead can the presence of a given target event be predicted from a sequence of events, before performance speed and accuracy are affected?

7.2. Method

7.2.1. Overview

After thorough consideration of various approaches, a number series test was selected to meet the requirements of this study, and thus reliably measure prediction in different language conditions. A computer-based experiment was developed in which participants were required to predict the occurrence of a number in a sequence of numbers that followed, by applying the observed logical pattern.

7.2.2. Participants

Participant were 40 Chinese–English bilingual non-aviation students (19 males and 21 females) enrolled in a study programme at Massey University, New Zealand, with Chinese as their native language (Mandarin dialect). The mean age of the participants was 24.60

years (*SD* = 4.72; *Range* = 18–42 years), and the mean duration of their stay in New Zealand was 2.67 years (*SD* = 3.57; *Range* = 2 months–21 years). Participants were assigned to one of the three English language proficiency groups (low, medium and high) according to the IELTS test scores they achieved when performing the test, which constituted a part of the experimental session. No participant reported any known hearing impairment.

7.2.3. Design

7.2.3.1. Overview

A 3 × 4 × 2 within-subjects experimental design was used to evaluate the effect of language switching on prediction. The three within-subjects factors were Language condition (Chinese, English or a Mix of both languages), Position (the position of a predicted number from the last number of a presented sequence; 1, 2, 3 or 4 steps ahead from the last heard number), and Pattern (the difficulty of a logical pattern that number sequences followed; assigned as 1 [easy] or 2 [hard] depending on the number of mathematical operations). A secondary aim was to test the effect of the between-subjects factor of language proficiency (low, medium and high) on performance in the English language and Mix conditions, for which one-way between-subjects ANOVA was used.

The benefits and potential shortcomings of using these designs were discussed in Study 2, namely an opportunity to manipulate more independent variables simultaneously and observe their effects on the dependent variable (within-subjects design), and an opportunity to compare the performance on the dependent variable (between-subjects design). This methodology is commonly used in aviation psychology research (e.g., Estival, Farris, & Molesworth, 2016; Barshi & Farris, 2013).

7.2.3.2. Development of the Experiment

Three language conditions were used to explore their effect on prediction: pure Chinese (L1), pure English language (L2), and a language switching condition (Mix) composed of English and Chinese stimuli. The order of the language conditions followed the rationale used in Studies 2 and 3 (e.g., Meuter & Allport, 1999); that is, monolingual language conditions did not follow each other. Half of the participants started the experiment with the L1 condition, and the rest started with the L2 condition.

In each language condition, participants were asked to predict whether a presented number continued a given sequence (yes or no). The numerical sequences followed a logical pattern, which was based on elementary arithmetic. Each language condition was developed to satisfy the following *criteria*: (i) the sequences were developed to assure that the performance would not be adversely affected by the difficulty of deducing a pattern, or solving an arithmetical operation, but would be attributable to the cognitive process of prediction; (ii) each of the sequences consisted of the same number of integers; (iii) the number sequences were developed following the same principle across the three language conditions to ensure that the difficulty of the conditions was balanced; and, (iv) the number to be predicted may have occurred in any one of four different positions after the last number of a sequence, and the positions were equally distributed within a language condition; that is, each appeared four times. This was designed to test Banbury et al.'s (2004) assumption that the further ahead in time operators predicted, the less accurate their predictions are likely to be.

The open-source application PsychoPy 1.82.01. (Peirce, 2007) was used for experiment development. The order of presentation of the stimuli in each language condition was set randomly to minimize the potential for any learning effects. The experiment was designed using the guidelines provided by the Massey University Ethics Committee, and peer review deemed it to be of low risk. A copy of the institutional low-risk notification can be found in Appendix F.

7.2.3.3. Measures

The dependent variables were the RT and the type and number of errors. Two types of errors were analysed, miss (i.e., a predicted number continued the sequence but participants responded that it would not) and false alarms (i.e., a predicted number did not continue a sequence of numbers but participants responded that it would).

Three within-subjects independent variables were investigated: Language condition, Position, and a logical Pattern. The Language condition factor had three levels; sequences were presented either in Chinese (L1) only, English (L2) only, or in a language switching condition (Mix), in which individual numbers within a single sequence were spoken in Chinese and English language. In each of the language conditions, the numbers in individual sequences were separated by fixed two-second intervals, owing to the findings of previous studies suggesting a small effect of the ISI factor. Individual number sequences (stimuli) were also separated by the two-second intervals.

The Position factor was developed according to Endsley (2000), who found that operators constrain the parts of a situation of interest based on how far away some element is, and how soon the element will have an impact on the task. Therefore, the Position factor, with four levels, referred to the position of a predicted number from the last number of the sequence. Position 1 meant that a predicted number (P) followed straight after the last number of a presented sequence (e.g., 2, 4, 6, 8, 10, **P**). This may be analogous to prediction of the immediate future. Position 2 referred to a position of a predicted number (P) continuing the sequence as the second after the last number of the sequence (e.g., 2, 4, 6, 8, 10, ?, **P**). Its prediction required applying the identified logical pattern twice to come to a decision. Predicted number on Position 3 continued the sequence as third after the last number of the presented sequence (e.g., 2, 4, 6, 8, 10, ?, ?, **P**). Finally, Position 4 meant that a predicted number was/or was not present four steps after the last number of the sequence (e.g., 2, 4, 6, 8, 10, ?, ?, ?, **P**). All four positions were chosen to be successive to explore how far ahead individuals were able to predict, and how prediction was affected by the increased distance of an event in the future.

The Pattern factor had two levels and it referred to a logical rule, which all of the number sequences followed. The logical rule was defined by the number of arithmetic operations, either one addition operation (Pattern 1), or two addition operations (Pattern 2). A number of studies have previously used the number of arithmetic operations to differentiate the complicacy and complexity of the tasks (e.g., McClain & Huang, 1982; Marsh & Maki, 1976). Pattern 1 consisted of a simple addition. There were two alternatives; each number in a sequence was either +2 or +3 greater than the previous number in that sequence (e.g., 3, 5, 7, 9, 11; or 3, 6, 9, 12, 15). Pattern 2 was more difficult as it consisted of two addition operations. The second number was +2 greater than the first number and the third number was +1 greater than the second number (e.g., 3, 5, 6, 8, 9). Alternatively, the second number was +1 greater than the first number and the third number was +2 greater than the second number (e.g., 3, 4, 6, 7, 9). The patterns were applied immediately from the first number of a sequence. The two levels of Pattern factor aimed to address prediction ability in situations of different difficulty (easy vs. hard).

Participants' ability in English formed a between-subjects independent variable, represented by the IELTS Listening test scores achieved when performing the test as a part of the experimental session.

7.2.4. Materials

7.2.4.1. Acoustic Stimuli

Participants were assessed on how quickly and accurately they could predict that a certain number, called the *predicted number,* would continue a given number sequence. Therefore, number sequences represented the acoustic stimuli of this study. Number sequences are given as finite sequences of numbers in certain patterns, which are solved by identifying the pattern and using it to continue the sequence to predict whether a number that was provided prior to the sequence (the predicted number), will appear or not somewhere in the sequence. Each sequence of five numbers was considered as one stimulus. All stimuli were spoken by a computerised female voice (OS X Text-to-speech programme) in Chinese (Mandarin) and English language, as in Studies 2 and 3, and recorded at the same speech rate. A propeller aircraft background noise was added to all stimuli, with a fixed SNR (its set up is described in the section 7.2.4.2).

To develop the sequences, three factors were considered. First, the effect of short-term memory on performance was reviewed to decide how many numbers should be included in a number sequence. The duration of short-term memory seems to be between 15 and 30 seconds (Atkinson & Shiffrin, 1971). The capacity of short-term memory is about 7 items, plus or minus 2 (Miller, 1956). Therefore, the sequences were developed as a reasonable trade-off between the requirement for a sufficient number of items for deducing a logical pattern and avoidance of the adverse effects of short-term memory capacity limitations. As a result, each of the sequences consisted of a maximum of five numbers and lasted approximately 13 s.

Second, the duration of the between-number intervals in the sequences was based on the known characteristics of short-term memory; an interval 2 s between numbers was selected. The rationale was in accordance with Peterson and Peterson's (1959) finding that the longer the delay of another stimulus, the less information was recalled. This meant that the interval had to be reasonably short to prevent loss of information—forgetting the just heard number—and thus losing the chance to identify the pattern. Additionally, the rationale was in accordance with the findings of Study 1 (see section 4.3.3), which revealed that the SA was more affected by bilingualism in the Terminal Control Area, where communication is frequent with very short intervals between the clearances.

Third, part of the stimuli development involved choosing a logical pattern that would meet the criteria; the sequences needed to be constructed so that the performance would not be adversely affected by the difficulty of solving an arithmetical operation. The Pattern factor had, therefore, two levels (easy vs. hard), depending on the number of mathematical operations. Pattern 1 was represented by one mathematical operation; that is, each number in a sequence was either +2 or +3 greater than the previous number in the sequence (e.g., 2, 4, 6, 8, 10; or 2, 5, 8, 11, 14). Pattern 2 was represented by two mathematical operations; that is, the second number in a sequence was +2 greater than the first number, and the following number was +1 greater than the second number (e.g., 2, 4, 5, 7, 8). Alternatively, the second number was +1 greater than the first number in a sequence, and the following number was +2 greater than the second number (e.g., 2, 3, 5, 6, 8). The two patterns each appeared eight times within each of the language conditions, and the order of presentation was randomised.

Each of the language conditions consisted of 16 sequences. The development of the 16 sequences in each of the language conditions was the fourth consideration; to create sequences that were different, yet equally difficult. Twelve of the sequences started with a different number in range the 0–11. The additional four sequences started with a number that had already been used at the start of a sequence, and a different logical pattern was applied to differentiate these sequences between each other. Moreover, the four sequence-beginning numbers that were used twice within a language condition, were different across the three language conditions so that no sequence was repeated twice across the three language conditions or within the condition. This allowed the construction of balanced language conditions.

The Mix condition was created so that the two languages, Chinese and English, alternated within each of the 16 sequences. In other words, the individual numbers in each of the number sequences were in alternating Chinese and English language; eight sequences began with a number spoken in Chinese and eight began with a number spoken in English. Subsequent use of numbers spoken in Chinese and English within a sequence were in a different order, except for two sequences, which had the same order of Chinese and English numbers. In these sequences, however, different numbers were used, and where possible, different patterns. Two sequences had the same pattern and order of Chinese versus English numbers, but the numbers themselves were different.

The predicted numbers were displayed visually for 2 s on a computer screen prior to a number sequence starting to play. The final consideration related to the stimuli was the format in which the predicted numbers should be displayed. According to Conrad (1964), visual information is encoded primarily acoustically; that is, visual information is translated into sounds. The predicted numbers were, therefore, displayed as Chinese characters in the L1 condition, and as Arabic numerals in the L2 condition. In the Mix condition, the predicted numbers were equally displayed in both Chinese characters and Arabic numerals, regardless of the language of the first number of a sequence. In other words, English number was equally preceded by a Chinese character or the Arabic numeral. The same logic was applied to sequences that began with a number spoken in Chinese language.

In summary, over the three experimental conditions, each participant was presented with 24 sequences that contained the predicted number and 24 that did not, giving 48 stimuli in total. Eight sequences contained a predicted number and eight sequences did not. In both cases, each of the four levels of Position factor was used twice—one with simple Pattern 1 and the other with Pattern 2.

In the study, participants were assessed on how quickly and accurately they could predict the presence of a certain number. In accordance with Marslen-Wilson’s (1985) finding that words can be processed before they are fully heard, participants were able to respond while a sequence was still playing—they did not have to wait until they had heard all five numbers of a sequence (stimulus). The RT (in seconds) was measured from the onset of the auditory stimulus; that is, from the first number of the sequence. The RT was adjusted using the subtraction method described in section 3.12, and this was used for the statistical analyses that followed. The stimuli list is included in Appendix G.

7.2.4.2. Speech to Noise Ratio

The same propeller aircraft in-flight interior noise was used as in Study 3. It was downloaded from the web page of Free Sound Effects FX Library (GRSites, n.d.) and added to the speech signal as background noise. The number sequences were all spoken by a computerised female voice set at the same level of volume. All stimuli were normalized to the same decibel level to add the desired SNR.

The SNR was fixed, meaning that it did not change across the stimuli and language conditions. Although the speech signal was adjusted by the gain effect to +10 dB more than the noise using Audacity 2.1.0, neither signal nor noise were amplified. Omitting the amplification effect in this study prevented the clipping distortion that adds the "microphone effect" in playback. This was corrected based on the findings of Study 3, when the noise was perceived by the participants as very high, yet not that high so any participants chose to discontinue the experimental session, as was explained they could do at the outset of the experiment. The two tracks (spoken number sequences and noise of propeller aircraft) were mixed to obtain the result of speech stimuli containing aircraft noise in the background.

7.2.4.3. English Language Proficiency

The English language proficiency of participants was evaluated using IELTS Listening Test 4. The free IELTS Listening test was downloaded from the public website (IELTS-up, n.d.). The test took approximately 30 minutes to complete and participants were provided an additional 10 minutes to transfer their answers to the answer sheet following the same instructions as given in the actual test.

7.2.5. Procedure

Participants were recruited from around the Massey University campus or by using social media between March 25 and June 31, 2017. During an appointment, the participants were introduced to the experimental situation, which consisted of the language switching experiment and the IELTS Listening test to measure English language proficiency in listening. The overall procedure was run in a single session and lasted approximately 1 hour (experiment and IELTS approximately 30 minutes each).

Participants started with the experiment to avoid any adverse effect of increased tiredness after performing the IELTS test. Participants first set up their own comfortable volume by following instructions for setting safe sound levels to protect their hearing. They were given time to practice and become familiar with the task. The experimental task required participants to predict the presence of a predicted number in a number series task. The predicted number was displayed visually on a computer screen for 2 s prior to the sequence of numbers being played. A card with each of the predicted numbers was also placed on the table in front of the participants by the experimenter to remind them of the number to be predicted. Next, participants were presented with spoken sequences containing five numbers that followed a certain logical pattern. Participants applied the recognized pattern to continue the sequence and decided whether the predicted number continued the sequence (*yes* response) or not (*no* response). Participants reacted by pressing '*yes*' or '*no*' on a keyboard, as soon as they had decided.

Unlike the previous study, participants were not able to ask for stimulus repetition. However, there was an option to press '*?*' on a keyboard when participants did not know the answer. Despite this option, participants were encouraged to try to predict and to take more time to decide. Simultaneously, they were instructed to respond as soon as they decided. Responses were possible even before participants heard the last number of the sequence. The experiment lasted approximately 25 minutes, depending on the speed of responses.

Following the experiment, participants completed IELTS Listening Test 4. The test took approximately 30 minutes. Participants were then assigned to one of three proficiency groups based on the following the IELTS categorization (British Council, n.d.b): Limited–

Competent users (*n* = 12; IELTS test scores: 3.5–6.0), Good users (*n* = 14; IELTS test scores: 6.5–7.0), and Very Good users (*n* = 14; IELTS test scores: 7.5–8.5).

Finally, participants were asked to provide demographic information on age, sex, the mean duration of their stay in New Zealand, and any known hearing impairment. The study was conducted anonymously, so participants were not asked to provide their names or contact details unless they wanted to receive a results summary. Participants were provided with a $10 voucher to the student dining hall, sponsored by the School of Aviation, as gratitude for their participation in the experiment.

A priori power analysis using G*Power software (Erdfelder, Faul, & Buchner, 1996) was used to determine that, with $\alpha = .05$, a total sample size of *n* = 28 would be sufficient for experimental power of .80, assuming an effect size of .25 (based on Ison, 2011) in a repeated-measures, within-subjects ANOVA. Additionally, *a priori* power analysis for a repeated-measures between-subjects design for four English language proficiency groups (based on the findings of Study 3) with $\alpha = .05$ and an effect size of $f = .40$ (based on Ison, 2011) suggested that a total of *n* = 30 participants would be sufficient for experimental power of .80.

7.3. Results

7.3.1. General Findings

As in the previous experiments, the data were screened for outliers, which were defined as z-scores greater than |3.29| as recommended by Tabachnick and Fidell (2007). This screening identified eleven outliers in the Language, Pattern, and Position data. Following the same rationale as in Study 2, after reviewing the effect of the outliers on the linear relationship between the data based on normal Q-Q plots, it was decided to exclude three outliers, and the rest were considered to be legitimate to retain, assuming that with a larger sample size they may represent the population. However, to ascertain that this would not affect the analysis, the data were also screened to verify that they met the assumptions required for the statistical tests. As in the previous analyses, the level of statistical significance, alpha, was set at .05 for all statistical tests, and all tests were conducted as two-tailed.

7.3.1.1. Prediction RT: The Challenge

As in Studies 2 and 3, consideration regarding RT was made prior to data analysis. A novel aspect was recognized, which was not relevant in the previous two studies, namely, the temporal aspect of prediction. Endsley (2000) recognized the temporal aspect of SA as critical and specified it as the need to know how much time is available until some event occurs or some action must be taken. Because of the temporal nature of prediction in a dynamic system, the measurement of RT became a greater challenge. The question arose as to how pure RT could be defined to allow accurate and representative measurement.

It was assumed that there was no benefit in understanding pure RT in the same way as in the previous two experiments; that is, as a subtraction of the stimuli durations from corresponding latencies. This method would not capture the process of prediction, which started before the actual response was made. The subtraction method, consequently, would not carry information about the time it took to make a prediction. However, using a mean RT, measured from the onset of a stimulus, would not provide sufficiently precise information about the prediction either, given that the pattern must first be recognized to make a prediction. The nature of number series stimuli indicated the need for correction. Adjustment of RT is not rare in experimental studies. For example, Zwitserlood (1998) measured RT from the onset of the altered word to a pseudoword in a spoken sentence. It can be assumed that the RT measurement depends on the goal of the measurement; that is, what is to be measured.

The understanding of RT as an *index of the complexity* of the inner process by which a result is accomplished (see section 3.12) led to the following correction method. Based on the SA definition (Endsley, 1995), it was assumed that to be able to predict, participants first needed to perceive the numbers (level 1 SA, recognition) and identify a pattern (level 2 SA; comprehend the way elements change over time) so that they could use it to predict the presence of a predicted number. Because two types of Pattern were used—simple and hard—it was assumed that prediction could begin after hearing at least two or three numbers of a sequence, depending on the Pattern. Therefore, the prediction time was measured from the onset of the third or fourth number of the spoken sequence, depending on the difficulty of the Pattern. Specifically, the pure RT that was used for the analyses was calculated by subtracting the duration of the first two or three digits of a sequence (depending on Pattern type) from the actual latency on that sequence. It should be noted that this RT is probably not precise either, but it can be considered as the most accurate approximation for the measurement purposes.

Final consideration revealed that because the language switching occurs within a sequence, and not between the sequences, it was not feasible to obtain switch costs and mixing costs. Nevertheless, general comparison of bilingual and monolingual conditions can suggest which of the conditions facilitated faster and more accurate responses. The mean durations of the stimuli and the mean RTs for each of the language conditions are presented in Table 27.

Table 27

Mean Response Times (RT), Mean Durations of Stimuli, and Mean Pure RT in Seconds in First (L1), Second (L2), and Language Switching (Mix) Conditions

	Language conditions					
	L1	*SD**	L2	*SD*	Mix	*SD*
Mean RT	15.333	4.364	14.696	3.245	14.759	3.856
Mean duration of stimuli	12.936	0.316	12.987	0.284	12.879	0.248
Mean pure RT	7.671	2.620	7.485	2.455	7.218	2.204

*SD = standard deviation

To provide a comprehensive analysis, and for clarity, the Results section is organised in line with the two main indices of performance; speed and accuracy. Each of these sections is further organised according to the within- and between-subjects variables. The Results section is completed by SDT measures.

7.3.2. Performance Speed

To explore the performance differences between three language conditions, a 3 × 4 × 2 within-subjects ANOVA was performed to test whether there was a difference in speed of performance attributable to different Language conditions (L1, L2 and Mix), Position of a predicted number (1, 2, 3, 4), and the difficulty of a logical Pattern (1, 2). As the repeated measures ANOVA can be sensitive to violations of sphericity (Schmider et al., 2010), Mauchly's test was considered and found violated. Therefore, degrees of freedom were corrected based on the values of Epsilon ($\varepsilon = .75$) (Girden, 1992).

The main effect of the Language condition factor was not found to be statistically significant ($p = .272$), but there was evidence of statistically significant main effects for the Position (Greenhouse-Geisser adjusted $F(1.554, 57.5) = 59.072$, $p < .001$, $\eta_p^2 = .615$), and Pattern factors (Huynh-Feldt adjusted $F(1, 37) = 4.137$, $p = .049$, $\eta_p^2 = .101$). The two-way interaction effect between Language condition and Position was found to be statistically significant (Greenhouse-Geisser adjusted $F(4.085, 151.128) = 3.095$, $p = .006$, $\eta_p^2 = .077$), as was the effect between Position and Pattern (Huynh-Feldt adjusted $F(2.655, 98.237) = 3.055$, $p = .031$, $\eta_p^2 = .076$). However, the two-way interaction effect between the Language condition and Pattern factors was not found to be statistically significant ($p = .659$). The three-way interaction between Language condition, Position and Pattern factors was statistically significant (Huynh-Feldt adjusted $F(5.729, 211.986) = 5.805$, $p < .001$, $\eta_p^2 = .136$). The *post hoc* analysis was conducted to determine only interactions related to the statistically significant main effects of Position and Pattern factors. The interaction plots are presented in Figure 9.

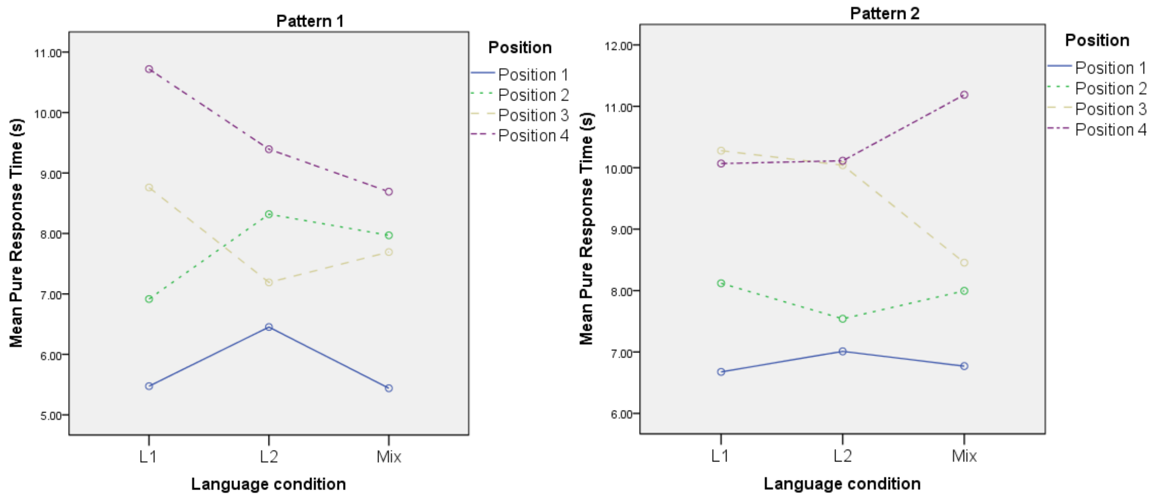


Figure 9. Interaction plots for Language condition, Position, and Pattern factors.

7.3.2.1. Performance Speed and Position

Pairwise comparison (Least Significant Difference[5]) of the four levels of the Position factor indicated that the further the predicted number was from the last number of the sequence, the longer it took to predict it (p < .001), for Position 1 (M = 6.304, SD = 2.513), Position 2 (M = 7.810, SD = 3.762), 3 (M = 8.736, SD = 3.727), and Position 4 (M = 10.029, SD = 5.094). Despite no statistically significant two-way interaction of Language condition and Position factors, it was considered relevant to provide the mean pure RTs across the three language conditions for the particular Position because of the main focus on the Language condition factor and consistency across the studies (see Table 28).

[5] Least Significant Difference was chosen for the *post hoc* test rather than Bonferroni, as the Bonferroni test controls for multiple comparisons, and thus is less sensitive to the significance between some of the groups. LSD does not make any adjustment for the number of comparisons. Using the Bonferroni correction could increase the risk of *type II* error (missing a difference that is present), which in the context of this study would be more important than *type I* error (identifying a difference that is not present). There was a relatively small number of comparisons (four for the Position factor and only two for the Pattern factor), and because of the novelty of the analysis, further investigation would be required prior to using it in a real-world environment.

Table 28

Mean Pure Response Times in Seconds and Standard Deviations (SD) across Language Condition and Position Factors

	Position 1		Position 2		Position 3		Position 4	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
L1	6.075	2.920	7.518	6.760	9.519	4.592	10.393	5.183
L2	6.731	2.449	7.930	3.131	8.616	4.326	9.754	4.219
Mix	6.105	3.026	7.983	3.239	8.073	3.057	9.939	7.382

7.3.2.2. Performance Speed and Pattern

The effect of difficulty of Pattern on speed of performance was measured by pairwise comparisons using the Least Significant Difference and indicated that performance was statistically significantly faster on Pattern 1 (*M* = 7.751, *SD* = 3.016) than on Pattern 2 (*M* = 8.688, *SD* = 3.914) (*p* = .049). Table 29 summarises the data for the two-way interactions of Language and Pattern factors.

Table 29

Mean Pure Response Times in Seconds and Standard Deviations (SD) across Language Condition and Pattern Factors

	Pattern 1		Pattern 2	
	M	*SD*	*M*	*SD*
L1	7.967	3.372	8.785	4.623
L2	7.839	3.149	8.676	4.206
Mix	7.448	2.947	8.602	4.036

7.3.2.3. Effect of English Language Proficiency on Performance Speed

Prior to testing whether performance speed differed between the three English language proficiency groups, the assumption of normality was evaluated and was found to be violated in the Mix condition as some of the group’s distributions were associated with skewness and kurtosis beyond the span of –2.0 and +2.0 (Cramer, 1998; George & Mallery, 2010). Owing to the violated assumption of normality in the Mix condition, a non-parametric Kruskal-Wallis H test was used, which indicated no statistically significant differences (*H*(2) = 1.653, *p* = .438).

To test the presence of differences between the English language proficiency groups in the L2 condition, and given the satisfied assumption of normality, a one-way ANOVA was conducted. However, there was no evidence of statistically significant differences (*F*(2, 37) = 1.125, *p* = .336). The descriptive statistics associated with participants' mean pure RTs across the three English language proficiency groups are reported in Table 30 and Figure 10. Note that five outliers are displayed in Figure 10. However, as explained in Study 3 (see section 6.3.2.5), these values were not defined as outliers.

Table 30

Mean Pure Response Times in Seconds across the English Language Proficiency Levels

	IELTS Listening	*n*	L2	*SD*	Mix	*SD*
Limited–Competent user	3.5–6.0	12	7.728	3.216	7.204	3.052
Good user	6.5–7.0	14	9.452	2.388	9.917	4.326
Very Good user	7.5–8.5	14	7.966	3.916	8.098	3.758

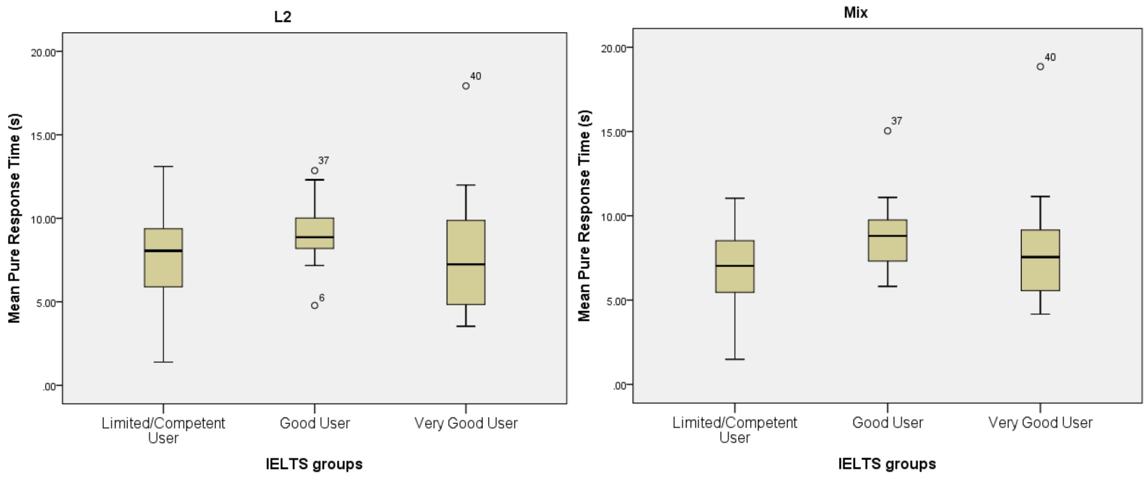


Figure 10. Profile plots for the English language proficiency groups in the L2 and the Mix conditions.

7.3.3. Performance Accuracy

7.3.3.1. Performance Accuracy and Language Conditions

Prior to testing whether performance accuracy differed between the language conditions, the assumption of normality was evaluated and was found to be violated as the distributions were associated with skewness and kurtosis beyond the span of –2.0 and +2.0 (Cramer, 1998; George & Mallery, 2010). Friedman's test indicated no statistically significant differences in performance accuracy between the language conditions (*p* = .374). Data are presented in Table 31.

Table 31

Error Types (Miss and False Alarm), Hits and Correct Rejections (CR), and Total Number of Errors across Language Conditions, and Percentage of Errors from 1920 Stimuli (%Error$_T$)

	Miss	False alarm	Hits	CR	Error totals	%Error$_T$
L1	37	21	283	299	58	
L2	40	34	280	286	74	
Mix	36	17	284	303	53	
Totals	113	72	847	888	185	9.63

7.3.3.2. Performance Accuracy and Position

Prior to testing whether performance accuracy differed across the levels of the Position factor, the assumption of normality was evaluated and was found to be violated as the distributions were associated with skewness and kurtosis beyond the span of –2.0 and +2.0 (Cramer, 1998; George & Mallery, 2010). Although there was evidence of statistically significant differences in performance accuracy between the four levels of this factor, based on Friedman's test ($\chi^2(3)$ = 16.144, *p* = .001), a Wilcoxon signed-rank test found the differences occurred only between Positions 2 and 3 (7.5% vs. 12.1%; *Z* = –2.440, *p* = .015, *r* = .39, risk ratio = 1.61). There was no evidence of a statistically significant difference between Positions 1 and 2 (*p* = .759), nor between Positions 3 and 4 (*p* = .967). The error counts across the four levels of the Position factor are presented in Table 32.

Table 32

Number (n_{errors}) and Percentage of Errors (%Error$_E$) across Levels of Position Factor and as a Proportion of Total Number of Stimuli in Each Position Factor Level (%Error$_P$)

	Position 1	Position 2	Position 3	Position 4
n_{errors}	32	36	58	59
%Error$_E$	17.30	19.46	31.35	31.89
%Error$_P$	6.67	7.50	12.08	12.29

Note. Total number of stimuli *N* = 1920

7.3.3.3. Performance Accuracy and Pattern

Prior to testing whether performance accuracy differed between the two levels of the Pattern factor, the assumption of normality was evaluated and was found to be violated as the distributions were associated with skewness and kurtosis beyond the span of –2.0 and +2.0 (Cramer, 1998; George & Mallery, 2010). Therefore, a non-parametric Wilcoxon signed-rank test was used, which indicated statistically significant differences in performance accuracy between Pattern 1 and 2 ($Z = –3.458$, $p = .001$, $r = .55$, risk ratio = 1.64). The risk of making an error on Pattern 2 was almost twice the risk of making an error on Pattern 1 (see Table 33).

Table 33

Number (n_{errors}) and Percentage of Errors (%Error$_E$) across Pattern Factor Levels and as a Proportion of Total Number of Stimuli in Each Pattern Factor Level (%Error$_P$)

	Pattern 1	Pattern 2
n_{errors}	70	115
%Error$_E$	37.84	62.16
%Error$_P$	7.29	11.98

Note. Total number of stimuli *N* = 1920

7.3.3.4. Effect of English Language Proficiency on Performance Accuracy

Prior to testing whether performance accuracy differed across the English language proficiency groups, the assumption of normality was evaluated and was found to be violated as the distributions were associated with skewness and kurtosis beyond the span of –2.0 and +2.0 (Cramer, 1998; George & Mallery, 2010). Therefore, a non-parametric Kruskal-Wallis H test was used, which revealed no difference in performance accuracy across the three English language proficiency groups ($H(2) = 1.450$, $p = .484$). Data are presented in Table 34.

Table 34

Number of Participants (n), Correct Responses and Errors, and Percentage of Errors (%Errors) across English Language Proficiency Levels

	IELTS Listening	*n*	Correct	Errors	%Errors
Limited–Competent user	3.5–6.0	12	507	69	11.98
Good user	6.5–7.0	14	605	67	9.97
Very Good user	7.5–8.5	14	623	49	7.29

7.3.4. SDT Measures

The SDT measures, sensitivity (*d'*) to distinguishing the signal from background noise and decision criterion (*C*) applied by participants to decide whether number would be present or not, were calculated according Stanislaw and Todorov’s (1999) recommendations. A summary of the data is presented in Table 35.

Table 35

Sensitivity (d') and Decision Criterion (C) of Prediction Task across the Three Language Conditions: Native (Chinese) Language (L1), Second (English) Language (L2), and Language Switching (Mix)

	z(HR)	*z(FAR)*	*d'*	*C*
L1	1.20	–1.51	2.71	0.155
L2	1.15	–1.24	2.39	0.045
Mix	1.23	–1.61	2.84	0.190

Findings indicated high detectability of stimuli from noise in all language conditions ($d' \geq 2.39$). In the Mix condition, participants were only slightly more prone to non-detection responses ($C = 0.190$) than in the L1 condition ($C = 0.155$), and were negligibly biased towards *no* responses ($C = 0.045$) in the L2 condition.

7.4. Discussion

The findings indicate that neither prediction performance speed nor accuracy differed between the language conditions (L1, L2, and Mix), which answers the first guiding research question. However, in terms of performance accuracy, there was evidence indicating greater miss rates than false alarm rates in the Mix condition. This was supported by further analysis of the SDT measures. This tendency towards *no* responses was only minor in the L2 condition and explains the greater false alarm rates in the L2 condition than in the Mix condition. In other words, participants were more prone to miss the presence of a predicted number when predicting it in the bilingual condition than in any of the monolingual conditions. Although this evidence is insufficient for making more general assumptions, it can indicate slightly better performance in monolingual conditions than bilingual, in terms of response bias.

Attention will be concentrated upon the performance speed, as it is important to explore reliability of the proposed approach to the response tome (RT) measurement. This discussion is parallel to, and concludes, the overall discussion of the subtraction method, which was based on the understanding of the RT either as an *index of complexity* of the inner process, or as an *index of achievement* (both explained in section 3.12). The RT decomposition method (as cited in Sternberg, 1969) has been used in previous studies (e.g., Geary et al., 1993; Widaman et al., 1989), however, it is unclear how the RT was measured. Therefore, there remains one unanswered question as to whether the findings would differ when the mean RT was used for analysis instead of the mean pure RT. Future research may provide some explanation to either support or refuse the suggested method of RT measurement.

Nonetheless, the analysis of performance speed revealed that the further the predicted number was from the last number of the provided sequence, the longer it took to predict it, consistent with Banbury et al.'s finding (2004). However, this could also be attributed to the calculation process; that is, the more steps required between the last number of the provided sequence and the predicted number, the longer it took to calculate, or apply the recognized pattern, to come to a decision. It appears that the odd–even rule (e.g., Krueger, 1986; Shepard, Kilpatric, & Cunningham, 1975) did not have a large facilitating effect on the speed of responses. Presumably, if participants had used the odd–even rule in this experiment, the responses would have been much faster. For example, when a predicted number was even, such as '22', and a sequence followed simple addition (+2), such as '2, 4, 6, 8, 10', it would be easy to predict the presence of this predicted number by continuing the sequence without calculation. However, the RTs suggested that this rule was not used.

In terms of accuracy of performance, a cut-off position of accurate prediction was found between Positions 2 and 3, where the number of errors increased. No difference in accuracy was observed between Positions 1 and 2, or between Positions 3 and 4. This provides the answer to the second guiding research question; predicting no more than two steps ahead is optimal, because after that, the number of errors increases. This appears to be somewhat contrary to the general rule, which states that the further ahead people are able to predict, the more time they have to take preventive action (Southwest EcoMotoring Club, n.d.). However, when considering the following two examples, the findings of this study and the general rule might not necessarily be in opposition.

The first example involves pilots listening to Automatic Terminal Information Service (ATIS) information[6], such as weather forecasts, which allows them to adjust their flight plan. The future condition is known (the weather), it is certain to some level (even though the weather can change) and will not change in response to what the pilot chooses to do; the pilot cannot affect the weather but can change how they are going to react. The second example illustrates the experimental situation tested in this study and involves an ATCO issuing climb instructions to one aircraft and extrapolating its trajectory to predict whether it will be on a potential collision course with some other aircraft flying at that level in future. The future condition, in this case, is unknown and uncertain, and is dependent on the

[6] ATIS broadcasts normally contain information on, for example, weather, air temperature, wind direction, visibility, type of approach(es) to be expected; the runway(s) in use; potential hazards, if any; significant runway surface conditions; altimeter settings, etc. (ICAO, 2001b)

ATCO's sequence of decisions. In other words, the future condition will be changed based on the ATCO's actions. With the ATCO's reactions, the future event can be changed at any time.

The weather conditions at the destination airport are independent from what the pilots experience during the flight; they occur in future, and therefore, pilots will have more time to prepare and decide their actions. However, for potential future hazards that depend on what is experienced now, the further ahead pilots try to predict, the longer it can take and the more erroneous the prediction may be. When pilots receive a weather forecast, they are aware of a condition in the future (event), yet they do not make predictions of the weather in the near future (prediction of an event) based on the pattern of weather events they are currently experiencing. They just prepare for the known future event. Whereas, in the ATC example, the ATCO predicts a potential future event (or hazard) based on the recognized pattern of a sequence of events that pilots, or the ATCO, have recently encountered and are encountering now. Therefore, it might be more appropriate to discuss preparation rather than prediction in the weather forecast example.

The findings related to the Pattern factor could be considered to indicate that the more complicated the logical pattern of a sequence of events, the longer it takes to predict. There were almost double the number of errors when applying the more complicated Pattern 2 for predictions, than when applying simple Pattern 1.

There was no evidence that the observed differences in performance speed or accuracy were attributable to the level of English language proficiency. Although no hypothesis was made regarding the effect of proficiency in English, this finding is somewhat surprising, as the task put greater demands on language skills than Studies 2 and 3 (Barshi & Farris, 2013). Further research would be needed to address this issue.

The total number of errors made in this study was almost 10%, suggesting that people are worse at predicting than at recognizing (Study 2) or comprehending (Study 3). This could also suggest task in Study 4 was more difficult than those in Studies 2 and 3. Contrarily, Jones and Endsley's (1996) investigation of SA errors using the ASRS database, indicated the opposite to be true in practice. Level 1 SA errors (recognition) accounted for 76.3% of all errors, whereas level 2 SA errors (comprehension) accounted only for 20.3%. The fewest errors were attributed to level 3 SA (prediction) (3.4%). Unfortunately, this study did not provide sufficient information to explain this contradiction. Further research would be necessary and, also, potentially significant for making implications for aviation safety.

When considering the challenges related to studying prediction, as described in section 7.1, two questions emerge. Can this contradicting result be explained, at least partially, by the extent to which errors of prediction can be identified based on language communication in real life? Are errors of recognition simply easier to detect than errors of comprehension and prediction? If there was a positive answer to the second question, the safety implications should be considered, because prediction errors can have more serious implications than people may be aware of. The point of being situationally aware is to avoid a hazard; to execute an appropriate reaction prior to a potential hazard occurring. Being aware of what is going on in the environment (CAA, 2014) suggests a somewhat passive process or a state of knowledge (Endsley, 1995), whereas when expressed as a prevention strategy, SA is a more dynamic, and proactive process. However, there may be other factors that could explain the difference between the findings of this study and those of Jones and Endsley (1996). Therefore, no assumptions are proposed at this stage of the research.

There are two potential limitations to the current study. As discussed previously, the study involved non-aviation participants; whether similar results would be obtained for the aviation population cannot be determined. However, the problems raised by this possibility are not unique to aviation psychology studies, as was explained in previous chapters. The study investigated fundamental cognitive processing rather than a specific realistic real-life operation. The final concern is related to the method of calculation of the pure RT. Few studies have used this method (e.g., Geary et al., 1993; Widaman et al., 1989) and its appropriateness should therefore be explored in further research.

In summary, the findings of the current study suggest that language conditions had no effect on the speed and accuracy of prediction performance in this study group. However, the findings suggested that the more complicated the pattern and the further into the future the predicted event is, the slower and less accurate the predictions will be. No specific implications were made at this stage of the research.

CHAPTER EIGHT

Study 5: Listening to Radio Calls over Background Talk

8.1. Introduction

To explain the context of this study, it is necessary to briefly recapitulate the results of Study 1, which revealed that approximately 67% of bilingual pilots and ATCOs communicate with their colleagues in their native [non-English] language when they are not communicating by radio. Such behaviour would be very natural. However, this requires them to switch attention from communication in their native language to English when making radio calls. For this thesis, the communication between colleagues when not communicating by radio will be termed '*background talk*'. Presumably, this background communication in between radio calls may cause some amount of distraction. When the radio calls are infrequent, such as during the cruise phase of flight, the conversation between pilots can be interrupted by an unexpected radio call requiring them to refocus their attention. A radio call—either addressed to their aircraft or not, depending on the call sign with which it begins—interrupts their conversation. The crew refocus on the ATCO's message and recognize their call sign (level 1 SA). The meaning of the message is understood (level 2 SA), leading to prediction of potential implications for the near future (level 3 SA).

A situation where non-native English-speaking pilots involved in conversation [presumably] in their native language perceive an unexpected radio call in English requires them not only to switch their attention but also to switch between languages. Consequently, SA can be impaired. However, it is unclear which of these two factors—the switch in attention or between the languages—accounts for a larger effect. Some errors and consequences can be related to attention being switched from background talk to unexpected radio calls. However, owing to the somewhat automatic priority given to ATC messages, pilots and ATCOs may have developed effective coping mechanisms, so that neither refocusing of attention nor language alternation pose a threat to effective operation. By investigating language switching between two different and parallel sources of speech (background talk and a radio call), the cognitive processes involved in the three levels of

SA were investigated under the condition of distraction; participants were distracted from their task by simultaneously listening to background talk.

Depending on the language of the background talk (native vs. English), and the language spoken on the radio in the airspace (monolingual English vs. bilingual), four situations can occur. First, when a crew speaks in their (non-English) native language (e.g., Chinese) and a radio message is issued in English, they need to switch between the languages as well as switching their attention. Second, if a Chinese-speaking crew is in Chinese airspace, they may perceive unexpected ATC messages in alternating Chinese and English languages. Third, it is possible that a crew communicates in English even though it is not the native language of either of the crew members, and then perceives the ATC messages in English. Fourth, a crew communicating in English as their second language operating in a bilingual air traffic environment may perceive the ATC messages in alternating English and Chinese languages. It is of interest to this study to investigate which of these four situations facilitates faster and more accurate responses. Both the speed and accuracy of responses can affect safe operations. For example, correct but late SA can impair safety as much as inaccurate but prompt SAs. However, little is known about performance differences between these four types of situations.

According to the CAA's *Flight-Crew Human Factors Handbook, CAP 737* (2014), a crew often lose SA because they are concentrating on other things or events. Distraction very often explains the loss of SA, yet it may not explain why the crew prioritised the way that they did (CAA, 2014). It can be challenging to investigate the development of SA based merely on verbal communications. Nevertheless, Hodgetts et al. (2005) found negative effects of party line communication on flight task performance, resulting in a greater deviation from the touchdown point on the runway, and associated self-reports of increased distraction and workload. Furthermore, an increase in flight checklist completion time was observed when background radio communication was present, and slightly more ATC calls were missed or queried in this condition. Hodgetts et al. (2005) suggested that background noise in the party line not only adds to pilot workload but may also impair cognitive performance on flight tasks. However, the effects of bilingual party line communication on performance have not yet been investigated. Three questions arose. If meaningful background talk was present, how would it affect the speed and accuracy of performance in a task? Is it harder to filter a message when the two independent streams of speech are in the same language or when they are in different languages? Does the dominance of either one of the two languages cause a difference in performance?

Selectiveness of attention might affect performance. For example, McAnally et al. (2010) summarised current knowledge in auditory change detection as follows. Listeners whose attention is focused on a certain sound stream commonly fail to detect unexpected changes in a concurrent stream. The changes to unattended features of a monitored sound stream can remain unnoticed, and when the listeners have not been instructed to attend to any target object in a sound stream, they performed poorly at detecting even substantial changes. This is somewhat contrary to the summary proposed by Demany et al. (2010), that change detection is to some extent automatic; that is, under certain conditions, participants could detect a change even if they did not pay attention to the relevant part of the sensory field, regardless of the sensory modalities (visual or auditory). Mattys et al. (2012) suggested that the effect of the cost of selective attention can be observed not only in the performance on the speech task, but also in the performance on another task. This can potentially explain the observations of Hodgetts et al. (2005), that flight task performance was also impaired by party line communication. Performance on a secondary task often declines when speech is more difficult to understand (Mattys et al., 2012). This redirects attention back to the question of which, if any, of the four background talk vs. language conditions in an ATC environment situations proposed in this section affect performance more than the others.

Although it tested selective attention within a visual rather than acoustic modality, Neisser and Becklen’s (1975) study is relevant here. Participants were presented with two different kinds of games on two video screens. One was a hand game and the second was a ball game. They were instructed to follow one game and ignore the other, and press a key when a significant, or odd event occurred. For example, in the ball game, one of the three male players was replaced by a female player. Participants performed this task without difficulty. However, when the two games were presented in the same fully overlapped visual field, they performed remarkably poorly. This suggests that attention must be focused on only one source of information to process it correctly. Whether the same effect would be found when two speech sources are presented simultaneously, such as the case of a pilot listening to background talk of a co-pilot and a radio call, cannot be determined. However, similar results were observed in the selective listening studies (Neisser & Becklen, 1975), in which two spoken messages were presented to a listener simultaneously.

According to Neisser and Becklen (1975), and Cherry (1953), a listener can report very little about the unattended message. Cherry's study (1953) revealed that participants noticed little of the unattended message, often not even realizing that at some point the language changed from English to German. This suggests that selectiveness of attention can affect not only level 2 SA, understanding, but also level 1 SA, recognition. For example, previous research on SA suggests that the amount of information individuals process depends on context and experience, with attention to less important information reduced when workload increases (for a review see Edgar et al., 2018). Consequently, a message can potentially be completely omitted (Kanarish, 2017); *"As pilots we almost all have a tendency to "tune out" non-essential transmissions."* (Survey: Participant 76). This raises the question of whether selectiveness of attention is fully controlled by the operator's will or choice, and what else might influence what is tuned out. According to Sumwalt and Watson (1995), it can be difficult for pilots to know where to focus their attention.

Neisser and Becklen (1975) suggested that words that are personally relevant or contextually meaningful may still be noticed even when they are unattended; a phenomenon known as the *'cocktail party effect'*. People are often able to notice that somebody spoke their name in a crowded hall while speaking with other people. This might suggest that pilots should be able to recognize their call sign—used to identify an aircraft—even though the radio was not attended to and they were involved in an informal conversation. However, Moray (1959) found that numbers were not recalled in a dichotic task; the only stimulus found to cause the cocktail party effect was the participants' own name. Whether a numeric call sign of an aircraft—which can be analogous to an aircraft own name yet comprised of numbers—would be recognized by pilots flying the aircraft while conversing with each other and not paying particular attention to the radio (e.g., while *en* route), has not yet been explored.

A final remark must be added on the discussion presented above. To explore selective attention, dichotic listening tasks (Hiscock, Inch, & Kinsbourne, 1999; Neisser & Becklen, 1975) were typically used. However, these tasks explore hemispheric lateralization. In a common dichotic listening task, participants wear headphones through which they are presented with different messages in each ear. These tasks differ from real-life aviation scenarios. Consequently, the results of dichotic listening tasks do not explain how switching attention between background talk and radio calls affect performance. Nor can they explain what effect the language of these two speech sources has on performance.

The guiding research question for Study 5 was:

Question 5: Does listening to two simultaneous messages affect performance on call sign recognition, error identification and prediction in monolingual and bilingual conditions?

8.2. Method

8.2.1. Overview

This study sought to investigate the effect of simultaneous listening to two sources of speech—background talk and an experimental task—on recognition (level 1 SA), error identification (level 2 SA), and prediction (level 3 SA) performance. Three computer-based tasks were developed as shortened and adapted versions of Study 2 (Call sign recognition task), Study 3 (Error identification task) and Study 4 (Prediction task). Therefore, the materials and design for Study 5 were similar to those of Studies 2, 3, and 4. The tasks were administered in consecutive order after general instruction in an introduction to the experimental session (see Figure 11).

General instruction
Call sign recognition with background talk
Error identification with background talk
Prediction with background talk

Figure 11. Study 5 flow diagram with three consecutive tasks performed under background talk; the Call sign recognition, Error identification, and Prediction tasks.

Participants were assigned to one of two experimental groups. One group was exposed to background talk in their native language (Chinese) while the other listened to background talk in their second language (English), and both groups simultaneously performed the three experimental tasks. In the first task, participants were asked to identify target call signs among distractors. The second task was to identify an error in simple arithmetic equations. The third task was to predict the occurrence of a particular number in a sequence of numbers by applying a logical pattern. More detailed explanation is provided in the following sections.

8.2.2. Participants

Forty-five Chinese–English bilingual non-aviation students (23 males and 22 females) enrolled in a study programme at Massey University, New Zealand, participated in this study. The mean age of the participants was 27.96 years (*SD* = 7.00; *Range* = 19–47 years) and Chinese was their native language (Mandarin dialect). The mean duration of their stay in New Zealand was 2.40 years (*SD* = 2.46; *Range* = 1 month–10 years). Participants were not assigned to any English language proficiency group because after data collection it was identified that the self-reported IELTS Listening test scores participants achieved before entering their study programme would not reflect their English language literacy. This was identified as limitation of the study. No participant reported any known hearing impairment.

8.2.3. Design

8.2.3.1. Overview

The primary aim of this study was to explore the effect of background talk on the Call sign recognition, Error identification and Prediction tasks, which were performed in two language conditions, either English only (L2) or a Mix of Chinese and English language stimuli. A mixed factorial design was used. The reason for the choice of the design throughout this thesis was to use the simplest analysis that can provide a valid answer to the research question. Within-subjects and between-subjects designs were employed separately in the previous studies. However, there was a difference between the designs of studies 2, 3, and 4, as compared to Study 5, regarding what was the critical comparison. Specifically, studies 2, 3 and 4 sought to compare RT across the three language conditions in order to find out which facilitates faster responses. Therefore, the choice was to employ within-subjects design, which allows detecting differences between related means. All participants were measured in all three language conditions. However, in Study 5, the aim was to explore the effect of Background talk factor. In other words, the aim was to investigate whether the presence of background talk either in the participants' native

(Chinese) or second (English) language affects the performance in any of the language conditions. Therefore, there were both within– and between-subjects variables. The Background talk factor was the fixed-effects factor (or between-subjects variable), which was constant across participants. The Language condition factor (L2 vs. Mix) was the random-effects factor (or within-subjects variable). Study 6 will explore whether the presence of background talk (yes vs no) affects the performance in any of the language conditions. The three tasks (Call sign recognition, Error identification, and Prediction) were shortened versions of those used in the previous three experimental studies, Studies 2, 3 and 4.

The main benefit of the mixed-design ANOVA is that it allows exploration of potential interactions between the Background talk and Language condition factors on the dependent variables, which were performance speed and accuracy. This allowed for more complex analysis of the four possible situations in which bilingual pilots or ATCOs may operate. The four situations were described in the Introduction section and were simulated by four testing conditions. The condition that facilitates faster and more accurate responses may be identified by comparing the four testing conditions on performance speed and accuracy.

The following four testing conditions were compared in this study, and were developed by the combination of the Background talk and the Language condition factors: (i) Participants performed each of the three tasks in a monolingual English condition (L2) while listening to background talk in their native, Chinese, language; (ii) Participants performed each of the three tasks in the Mix condition of alternating Chinese and English stimuli while simultaneously listening to background talk in Chinese; (iii) Participants performed each of the three tasks in the L2 condition while listening to background talk in their second language, English; and, (iv) Participants performed each of the three tasks in the Mix condition while listening to background talk in English.

Although all participants could have completed all conditions, this was deemed to significantly increase workload and duration of the experiment. Therefore, it was decided that each of the participants would perform each of the three tasks in the L2 and the Mix condition while simultaneously listening to a background talk only in one language, either English or Chinese. Thus, the background talk was a fixed-effects factor (or between-subjects variable).

The potential shortcomings of using this design were considered in the previous studies and related to the general limitations of the experimental design. The methodology is commonly used in aviation psychology research (e.g., Barshi & Farris, 2013; Estival, Farris & Molesworth, 2016).

8.2.3.2. Development of the Experiment

To explore the effect of background talk in two different languages while simultaneously performing tasks, three issues were considered: the experimental tasks, the language conditions, and the background talks.

To explore the effect of background talk on SA, participants performed all three tasks of Call sign recognition, Error identification and Prediction, which were abbreviated versions of the tasks used in Studies 2, 3, and 4. Consequently, the difficulty of the entire experiment increased for participants. Therefore, some adjustments to these tasks were made: (i) variables that were not found to be statistically significant in previous corresponding studies were excluded, and variables found to have statistically significant effects were considered individually, to develop a reasonable number of measures; (ii) the number of stimuli in each of the three tasks were reduced to maintain an acceptable workload for the participants; and, (iii) noise used for the stimuli development was considered separately and will be described in more detail in section 8.2.4.1.4.

Because of the increased difficulty of the experiment when performing three consecutive tasks while listening to background talk, the native language condition (Chinese; L1) was not used. It was excluded to limit the workload of the study, and because in real life such a condition seldom occurs, especially at international airports. Therefore, two language conditions were developed. These were analogous to the two possible real-life scenarios, which are the monolingual English and bilingual air traffic environments. Participants performed each of the three tasks (Call sign recognition, Error identification, and Prediction tasks) in two language conditions: pure English language (L2), and a language switching condition (Mix) composed of English and Chinese stimuli. The order of the language conditions was counterbalanced.

All three tasks were performed while listening to background talk. To develop thorough experimental methodology, the following *criteria* for the use of background talk were formulated: (i) the language of the background talks had to be the same language used for experimental language conditions; participants' native (Chinese) and second language (English) were used; (ii) the content of the background talks should be similar across the two Background talk factors, due to the potential effect of the content of background talks on listening attention of participants; (iii) the theme of background talk content should be neutral, somewhat interesting, and spoken in general conversational style, and, (iv) recorded using the same computerised voices as for the stimuli in the experimental tasks.

To decide how participants would be assigned to one of the two levels of the Background talk factor—Chinese or English language—various random assignments methods were reviewed (e.g., Smith, Morrow, & Ross, 2015; Suresh, 2011; Viera & Bangdiwala, 2007) with the primary aim of avoiding allocation bias. The following rule was applied, which was based on alternation of the two levels of the Background talk factor and participants' gender: Participants were pre-randomized into a group based on their gender and order of participation; the first participant was assigned to the Chinese background talk group, and the next participant of the same gender was assigned to the alternate, English, background talk group. Because the first participant was female and listened to background talk in Chinese, the next female participant was automatically assigned into a group listening to background talk in English. This method ensured a balanced gender distribution within the groups and was considered easy and not biased by examiner or participant choice. However, the method requires approximately equal numbers of male and female participants, a condition which Study 5 met.

In summary, 22 participants listened to background talk in English while performing the experiment (consisting of three tasks) and 23 participants listened to background talk in Chinese while performing the same three computer-based tasks. Both groups performed the same three consecutive tasks in the same order, starting with the easiest task of Call sign recognition (level 1 SA), continuing with the Error identification tasks (level 2 SA) and finishing with the Prediction task (level 3 SA). All three tasks were run in two language conditions, L2 and Mix, and the language conditions were counterbalanced.

The open-source application PsychoPy 1.82.01. (Peirce, 2007) was used for the experiment development. The order of presentation of the stimuli in each language condition was set randomly to minimize the potential for any learning effects. The experiment was designed using the guidelines provided by the Massey University Ethics Committee, and peer review deemed it to be low risk. A copy of the institutional low-risk notification can be found in Appendix H.

8.2.3.3. Measures

The dependent variables were the RT and the type (miss and false alarms) and number of errors.

The random-effects factor, or within-subjects factor, was the Language condition consisting of two levels; English only (L2) and the Mix condition of Chinese and English language stimuli. The presentation order of the language conditions was randomized.

In the Prediction task, there was one additional within-subjects variable, Position factor, which refers to the position of the number that was predicted after the last number of a sequence. The Position factor was retained because of its significance in Study 4 and its explanatory significance for prediction as the third level of SA. However, the Position factor had only two levels (Position 1 and 2), to keep the workload in tolerable limits. After identification and extrapolation of a logical pattern between the numbers in a sequence, participants predicted whether a particular number, called the predicted number, would or would not come immediately after the last number of the sequence (Position 1) or in the second position after the last number of the sequence (Position 2).

The fixed-effects factor, or between-subjects factor, was the Background talk factor (Chinese vs. English). The effect of English language proficiency was not examined, because the self-reported English language listening proficiency test scores participants achieved before entering a study programme were incomplete.

8.2.4. Materials

8.2.4.1. Acoustic Stimuli

Participants were assessed on how quickly and accurately they could recognize a target call sign (Task 1), identify an error in simple arithmetic equations (Task 2), and predict whether a certain number continued given number sequence (Task 3). Therefore, three sets of acoustic stimuli were developed. These were similar to those used in Studies 2, 3, and 4. A more detailed description of the stimuli for each corresponding task will be given in the following sections. This section discusses the common characteristics of the stimuli across the tasks.

All stimuli were spoken by a computerised female voice (OS X Text-to-speech programme) and recorded over a propeller aircraft background noise (using Audacity 2.1.0) with a fixed SNR (SNR set up is described in section 8.2.4.1.4). The presentation of stimuli was randomized. The randomization of stimuli in the Mix condition had potential benefits and shortcomings. The primary benefit was that it simulated real-world communication on a radio in which transmissions in two languages follow an unpredictable pattern, rather than a precise regularly alternating order of one ATC message in English followed by a message in Chinese (e.g., Chinese–English–Chinese–English etc.). Unpredictability, however, might also be a potential shortcoming. Because of the unpredictable sequence of stimuli, it was assumed that there might be periods in which several stimuli in one language could follow one after another in the language switching (Mix) condition (e.g., Chinese–Chinese–English–Chinese etc.). This was not considered a limitation, given the unpredictability of the language in which an ATC message is uttered in a bilingual air traffic environment. Creating conditions analogous to real-world operations was the primary goal.

Although the stimuli in all three tasks were separated by 3 ISIs (1 s, 4 s and 9 s), it was not considered as a within-subjects variable because its effects in Studies 2, 3 and 4 were predominantly not statistically significant. The primary purpose of the ISIs was to allow listening to background talk.

The on-line method of making a response, which was described in section 3.12, was used in this study; participants could respond while a stimulus was playing. This required similar consideration of the RT measurement; that is, the subtraction method was used and will be explained in corresponding sections.

8.2.4.1.1. Stimuli: Call Sign Recognition Task

Stimuli for the Call sign recognition task were developed following the same procedure as described in section 5.2.4.1, except for the Similarity factor, which was excluded from the analysis. That is, three-digit numbers were used to represent call signs. The target number remained the same within each language condition but differed across the two conditions (L2 and Mix). Thus, there were two different targets, each presented five times in a corresponding language condition. In the Mix condition, the target stimulus was presented in both languages; two in Chinese and three in English, based on the findings of Study 1 that, in real life, the use of English language dominates over the radio frequency (see section 4.3.2). The target stimuli were randomly distributed among 10 distracting stimuli, giving 15 stimuli in each of the two experimental language conditions. The probability of a distractor being presented was 67%, and the probability of a target was 33%. Over the two experimental language conditions, each participant was, therefore, presented with 20 distracting stimuli and 10 target stimuli, giving a total of 30 stimuli. All stimuli were different from each other. The stimuli list can be found in Appendix I.

8.2.4.1.2. Stimuli: Error Identification Task

Stimuli for the task of Error identification were developed following the same procedure as described in section 6.2.4.1. In the two language conditions, L2 and Mix, participants were asked to identify a mistake in simple arithmetic equations (additions and subtractions). Each of the two language conditions consisted of 12 stimuli out of which four were incorrect and eight were correct. The probability of an incorrect equation being presented was approximately 33%, and the probability of a correct equation being presented was 67%. The goal was to simulate the infrequent presence of errors in aviation communications and test the effect of background talk on error identification.

Over the two experimental language conditions, each participant was therefore presented with 16 correct (eight additions and eight subtractions) and eight incorrect equations (four additions and four subtractions), giving 24 stimuli in total. To create a balanced Mix condition, correct and incorrect equations were presented in English and Chinese language; that is, two incorrect equations were in English and two in Chinese, and, of the eight correct equations, four were recorded in English and four in Chinese. All stimuli were different from each other. The stimuli list can be found in Appendix I.

8.2.4.1.3. Stimuli: Prediction Task

In the third task, participants were asked to predict whether a number they were shown on a computer screen prior to hearing a number sequence would follow the logical pattern to be present somewhere in that sequence (*yes* response) or not (*no* response). The acoustic stimuli were designed following the same criteria as in Study 4. Several adjustments were made to ensure reasonable workload.

The sequences across the two language conditions were created to be of equal complexity. Each of the language conditions (L2 and Mix) contained eight sequences, with each sequence consisting of five numbers. Thus, participants responded to 16 sequences in total. In the L2 condition, all numbers in the sequences were recorded in English. In the Mix condition, each of the sequences consisted of numbers in alternating English and Chinese language. Of the eight sequences in the Mix condition, half consisted of three numbers spoken in English and two numbers in Chinese, and half consisted of two numbers spoken in English and three in Chinese. Next, three sequences started with a number in English, and five started with a number spoken in Chinese. Subsequent alternation of Chinese and English spoken numbers within the sequences was random. The order of languages in which the numbers within a sequence were spoken was different across the sequences.

Intervals between the numbers in a sequence differed from the 2 s intervals used in Study 4. Numbers in each of the sequences were separated by three intervals, 1 s, 4 s and 9 s, to allow for listening to background talk. Because one sequence was considered a stimulus, the three intervals between the numbers in a sequence were set up during the development phase. Four combinations of the three intervals were developed to achieve random distribution. Each of the four combinations was used twice. The numbers in the sequence could be separated by the following four combinations of intervals starting after the first number of a sequence: (i) 4 s–9 s–1 s–4 s; (ii) 9 s–4 s–1 s–4 s; (iii) 1 s–9 s–4 s–4 s; and (iv) 1 s–4 s–9 s–1 s.

In each of the number sequences two simple logical patterns were applied; every subsequent number in a sequence was increased either by 2 or by 3 (e.g., 2, 4, 6, 8, 10, or 2, 5, 8, 11, 14). The logical pattern was not included as a within-subjects variable in the analysis, given that the findings of Study 4 revealed an effect of the more complicated Pattern 2. The simple pattern was used to develop number sequences. Unlike pattern, Position was considered a within-subjects variable. It had two levels, Positions 1 and 2. This was to investigate how many steps ahead participants were able to predict while simultaneously listening to background talk.

Next, the predicted number was displayed on a computer screen before a sequence started to play. In contrast to Study 4, predicted numbers were presented visually in numerical, Arabic form regardless of the language condition (L2 or Mix), with both Chinese and English spoken numbers. Chinese characters were not used. The onset of the sequences was set to be 2 s after the presentation of the predicted number on the screen. The predicted number was also placed on a small card in front of participants, to prevent any confounding effect of short-term memory limitations on performance, as was discussed in Study 4.

In the L2 condition, the predicted number was present four times (*yes* response) and absent four times (*no* response). In the Mix condition, the predicted number was present three times (*yes* response) and absent five times (*no* response). This was caused by a programming error at PsychoPy, and was identified as a limitation of the study. The stimuli list can be found in Appendix I.

8.2.4.1.4. Speech to Noise Ratio

The same propeller aircraft in-flight noise (GRSites, n.d.) was used as in Studies 3 and 4 and added to the speech signal as background noise. As before, the SNR was fixed. The speech signal was adjusted by the gain effect to +10 dB more than the noise using the Audacity 2.1.0. programme. Neither signal nor noise were amplified.

8.2.4.2. Background Talk

Background talk that met the criteria proposed during the development phase of the experiment (see section 8.2.3.2) was sourced from the Technology, Entertainment, Design (TED) talks. TED is a non-profit organization devoted to spreading ideas. It began in 1984 as a conference, and today covers almost all topics, usually in the form of short talks (TED Conferences LLC, n.d.). The advantage is that talks contain transcripts in various languages, including English and Chinese.

This allowed choosing simple but interesting stories, using their official transcript in Chinese and English language and recording them using the OS X Text-to-speech computer programme. Although the original TED talks selected for the research were in English, the speakers had different accents, which was identified as a potential confounding variable. To standardize the talks in English and Chinese, both transcripts were recorded by the same computerised female voices that were used for recording the acoustic stimuli in the previous studies. No background noise was added to the talks. Four stories were recorded, but only two were required in majority of the experimental sessions. In only a few cases was a third background talk needed; because of a participant's response latencies in the experimental tasks. Transcripts of the background talks can be found in Appendix J.

8.2.4.3. English Language Proficiency

Participants reported their IELTS Listening test scores that they achieved for admission to a Massey University study programme, which was similar to Study 2. The reason their ability was not actually tested in this study was based on the complicacy and complexity of this study and the time demands of the IELTS Listening test; that is, testing English language ability would add too much time to that required of participants. Moreover, measurement of the participants' English language proficiency was not the primary aim of this study. As such, this was considered to be sufficient information, given the observed (minimal) effect of English language proficiency on performance in the previous studies. However, after data collection was completed, it was found that the self-reported IELTS test scores did not reflect the actual English language literacy and, therefore, they were not used for the analysis. The reasons will be described in more detail in the Discussion, as this was identified as a limitation of the study.

8.2.5. Procedure

Participants were recruited from around the Massey University main campus between August 19 and September 30, 2017. They were assigned to a background talk condition—English or Chinese language—using the allocation method described in section 8.2.3.2. Subsequently, they were introduced to the experimental situation. Two computers were used to generate two simultaneous sources of speech. The three experimental tasks were run on one computer, with a screen and speakers located in front of participants. The second source of speech, background talk, was located to the left of the participants. Participants first set up their own comfortable volume by following instructions for setting safe sound levels to protect their hearing. The background talk was set at approximately the same volume as the experimental tasks. Each of the three experimental tasks had a practice trial. Participants were given time to practice and fully comprehend the task. All tasks were run in a single experimental session lasting approximately 40 minutes, depending on the speed of responses.

While listening to the background talk, participants were asked to perform three computer-based tasks. The procedure and instructions for each of the tasks were identical to those of Studies 2, 3 and 4. In summary, participants were first asked to press “Yes” on a keyboard when they heard a target number, and “No” for a distracting stimulus. Next, the task was to press “Yes” when the presented equation was correct and “No” when it was incorrect. Lastly, participants were instructed to press “Yes” when a predicted number continued the following sequence, and “No” when the predicted number was not present in the sequence. Participants could not replay the stimuli, nor was there an option to skip the response when they did not know.

The background talks started to play as soon as the experimental tasks, including the practice trials, began. To allow participants to fully concentrate, comprehend the task or ask questions, no background talk was played while they read the instructions after each language condition for a corresponding task, participants were asked to write down all the information they could remember from the talk they had just heard. This ensured that participants really paid attention to the background talk, as the primary interest of the study was to test its effect on task performance. A sheet with printed instructions (*"List everything you can remember from the talk you just heard"*) was provided to the participants. They wrote down the answer, whether it was just a single word or entire sentences. This was repeated six times; after each language condition of each of the three tasks.

After the experiment, participants were asked to provide demographic information on age, sex, the mean duration of their stay in New Zealand, and the IELTS Listening test scores they obtained when entering a Massey University study programme. Participants were given $5 voucher to a supermarket, sponsored by the School of Aviation, for participating in the experiment.

The study was conducted anonymously, so participants were not asked to provide any identification details, unless they wanted to receive a results summary. *A priori* power analysis using G*Power software (Erdfelder, Faul, & Buchner, 1996) was used to determine that with $\alpha = .05$, a total sample size of $n = 34$ would be sufficient for an experimental power of .80, assuming an effect size of $f = .40$ (based on Ison, 2011), for two groups of background talk language and four measurements, for a between-subjects ANOVA.

8.3. Results

8.3.1. General Findings

All data were screened for outliers, which were defined as z-scores greater than |3.29| as recommended by Tabachnick and Fidell (2007). The screening revealed one outlier in the Call sign recognition task, one in the Error identification task, and no outliers in the Prediction task. These were excluded following the same rationale and procedure as described in Study 2. The data were also screened to verify that they met the assumptions required for the statistical tests. As in the previous analyses, the level of statistical significance, alpha, was set at .05 for all statistical tests, and all tests were conducted as two-tailed.

Considerations about the measurement of RT in each of the three tasks followed the same rationale as in the corresponding studies 2, 3 and 4, and the same procedure was chosen. Generally, mean pure RT was calculated as a subtraction of the duration of a stimulus from the RT on that stimulus. The mean pure RT of the third task was computed by subtracting the duration of the first two digits of each sequence from the RT on that sequence, because only simple patterns were used, and therefore, the minimum number of digits that needed to be heard to identify a pattern was two. The mean pure RT together with the mean durations of the stimuli for all three tasks are presented in Table 36.

Table 36

Mean Response Times (RT), Mean Durations of Stimuli, and Mean Pure RT in Seconds in Second (L2) and Language Switching (Mix) Conditions across the Three Tasks with Background Talk (Call Sign Recognition, Error Identification, and Prediction)

	Call sign recognition with background talk			
	L2	*SD**	Mix	*SD*
Mean RT	2.152	0.499	2.219	0.637
Mean duration of stimuli	1.568	0.068	1.388	0.276
Mean pure RT	0.586	0.499	0.816	0.637
	Error identification with background talk			
	L2	*SD*	Mix	*SD*
Mean RT	3.759	0.799	3.318	0.835
Mean duration of stimuli	2.065	0.083	1.922	0.148
Mean pure RT	1.694	0.799	1.396	0.835
	Prediction with background talk			
	L2	*SD*	Mix	*SD*
Mean RT	12.375	4.268	11.457	3.907
Mean duration of stimuli	20.749	1.493	20.422	1.235
Mean pure RT	5.497	4.268	4.657	3.907

*SD = standard deviation

To provide a comprehensive analysis, and for clarity, the Results section is organised in line with the two main indices of performance, speed and accuracy, for each of the three tasks. SDT measures are provided at the end of the Results section.

8.3.2. Performance Speed

8.3.2.1. Call Sign Recognition Task

A mixed-design 2 × 2 ANOVA with Language condition (L2 vs. Mix) as a within-subjects factor and Background talk (English vs. Chinese) as a between-subjects factor was used to compare performance speed of call sign recognition. Because the within-subjects variable had only two levels, and there was therefore only one set of difference scores for the comparison of variance, Mauchly's test was not considered (Field, 2017).

The main effect of the Language condition factor was found to be statistically significant ($F(1, 42) = 5.741$, $p = .021$, $\eta_p^2 = .120$), with participants performing faster in the L2 condition ($M = 0.593$, $SD = 0.501$) than in the Mix condition ($M = 0.754$, $SD = 0.499$). There was no evidence of a statistically significant main effect of the Background talk factor ($p = .554$), or evidence of any significant interaction effect of the Language condition and Background talk factors ($F(1, 42) = 0.005$, $p = .644$, $\eta_p^2 = .000$). Table 37 shows the mean pure RT in seconds for each of the language conditions and background talks.

Table 37

Mean Pure Response Times in Seconds across Language Condition and Background Talk Factors

		Language condition			
		L2	*SD*	Mix	*SD*
Background talk	Chinese	0.632	.486	0.798	.540
	English	0.555	.527	0.711	.461

8.3.2.1.1. Switch Costs and Mixing Costs

To test the presence of switch and mixing costs, an analysis of the sequence of alternating English and Chinese stimuli was conducted. Prior to testing whether the differences were statistically significant, the assumption of normally distributed differences between pairs was examined and was found to be violated, as the skewness and kurtosis levels were beyond the span of –2.0 and +2.0 (Cramer, 1998; George & Mallery, 2010). Therefore, a non-parametric Wilcoxon signed-rank test was conducted, which indicated neither switch costs (*p* = .382) nor mixing costs (*p* = .070) were statistically significant (see Table 38).

Table 38

Median (Mdn) Pure Response Times in Seconds on Chinese and English Word Stimuli in Monolingual (L2) and Language Switching (Mix) Conditions; Mixing Costs and Switch Costs

L2		Mix			Switch costs
Sequence	*Mdn*	Sequence	*Mdn*	Mixing costs	(L2 vs. Mix)
L2 → L2	.598	L2 → L1	.605	0.085	→ L2: 0.092
		L1 → L2	.690		

8.3.2.2. Error Identification Task

To explore the effect of the Background talk factor (Chinese vs. English) on performance of error identification in monolingual (L2) or bilingual (Mix) conditions, a mixed-design ANOVA was conducted. It indicated a statistically significant main effect of the Language condition factor ($F(1, 41) = 15.341$, $p < .001$, $\eta_p^2 = .272$). The error identification performance was faster in the Mix condition (*M* = 1.299, *SD* = 0.565) than in the L2 condition (*M* = 1.684, *SD* = 0.807). Mean data are presented in Table 39. There was no evidence of a statistically significant main effect of the Background talk factor (*p* = .163) or any evidence of a significant interaction effect of the Language condition and Background talk factors ($F(1, 35) = 0.056$, $p = .814$, $\eta_p^2 = .001$).

Table 39

Mean Pure Response Times in Seconds and Standard Deviations (SD) across Language Condition and Background Talk Factors

		Language condition			
		L2	*SD*	Mix	*SD*
Background talk	Chinese	1.828	.861	1.420	.679
	English	1.539	.740	1.178	.394

8.3.2.2.1. Switch Costs and Mixing Costs

To test the presence of switch and mixing costs, the sequence of alternating English and Chinese stimuli was analysed. The assumption of normally distributed differences between pairs was examined and was found to be violated, as the skewness and kurtosis levels were beyond the span of –2.0 and +2.0 (Cramer, 1998; George & Mallery, 2010). A non-parametric Wilcoxon signed-rank test indicated statistically significant mixing costs, in particular, the responses were faster when switching to a Chinese stimulus after hearing a stimulus in English (*Mdn* = 0.780) in the Mix condition than when switching to an English stimulus after exposure to a Chinese stimulus (*Mdn* = 1.285) (*Z* = –5.527, *p* < .001, *r* = .481). However, there was no evidence of statistically significant switch costs (*p* = .086). Median pure RTs are presented in Table 40.

Table 40

Median (Mdn) Pure Response Times in Seconds on Chinese and English Word Stimuli in Monolingual (L2) and Language Switching (Mix) Conditions; Mixing Costs and Switch Costs

L2		Mix			Switch costs
Sequence	*Mdn*	Sequence	*Mdn*	Mixing costs	(L2 vs. Mix)
L2 → L2	1.506	L2 → L1	0.780	0.505	→ L2: 0.221
		L1 → L2	1.285		

8.3.2.3. Prediction Task

A mixed-design 2 × 2 ANOVA was conducted to compare prediction performance between the two language conditions (L2, Mix) as a within-subjects factor while simultaneously listening to background talk (English vs. Chinese), which was the between-subjects factor. There was evidence of a statistically significant main effect of Language condition factor ($F(1, 43) = 5.286$, $p = .026$, $\eta_p^2 = .109$), and Position factor ($F(1, 43) = 8.905$, $p = .005$, $\eta_p^2 = .172$), suggesting faster responses in the Mix condition ($M = 4.644$, $SD = 4.036$) than in the L2 condition ($M = 5.489$, $SD = 4.477$). The responses were found to be faster when a predicted number occurred fewer steps ahead, at Position 1 ($M = 4.664$, $SD = 4.4038$) rather than Position 2 ($M = 5.469$, $SD = 4.475$).

The two-way interaction effect of Language condition and Position factors was not found to be statistically significant ($p = .965$). The three-way interaction effect of Language condition, Position and Background talk factors ($p = .327$) was not significant. There appears to be no evidence of a statistically significant main effect of Background talk factor ($p = .436$). Data for the Language condition and Background talk are presented in Table 41.

Table 41

Mean Pure Response Times in Seconds and Standard Deviations (SD) across Language Condition and Background Talk Factors

		Language condition			
		L2	*SD*	Mix	*SD*
Background talk	Chinese	5.830	4.595	5.224	4.101
	English	5.149	3.976	4.064	3.694

8.3.3. Performance Accuracy

Table 42 summarises the error counts and rates for the three tasks performed when listening to Chinese or English background talks. Prior to examining which of the background talks had a greater adverse effect on performance accuracy, the assumption of normality was evaluated and was found to be violated as the distributions were associated with skewness and kurtosis beyond the span of –2.0 and +2.0 (Cramer, 1998; George & Mallery, 2010).

A non-parametric Wilcoxon signed-rank test was conducted, which indicated that the differences in performance accuracy between Chinese and English background talks were not statistically significant on any of the three tasks; Call sign recognition (*p* = .602), Error identification (*p* = .457), and Prediction (*p* = .925).

Table 42

Number and Percentage of Errors across the Tasks and Background Talks

	Chinese	%	English	%
Call sign recognition	53 (690)	7.681	48 (660)	7.273
Error identification	90 (552)	16.304	72 (528)	13.636
Prediction	50 (368)	13.587	49 (352)	13.920
Totals	193		169	

Note. Total number of stimuli within each of the tasks and background talks are presented in parentheses. Of the participants, 22 conducted the tasks with English background talk and 23 with Chinese background talk.

Table 43 summarises the error counts and rates for the three tasks performed in the L2 and Mix conditions. A Wilcoxon signed-rank test was conducted to compare performance accuracy across the language conditions. The risk of making an error in the Mix condition was almost five times the risk of making an error in the L2 condition in the Call sign recognition task while listening to a background talk (*Z* = –4.654, *p* < .001, *r* = .69, risk ratio = 4.61). There was no evidence of a statistically significant difference in performance accuracy between the L2 and Mix conditions on the second task, Error identification (*p* = .309). In the third task, Prediction, in the L2 condition, the risk of making an error was almost two times the risk of making an error in the Mix condition (*Z* = –2.574, *p* = .010, *r* = .38, risk ratio = 1.58).

Table 43

Number and Percentage of Errors across the Tasks and Language Conditions, Total Number of Errors (Totals), Total Number of Stimuli within Tasks (N), and Percentage of Total Number of Errors (%ErrorT) within Tasks

	L2	Mix	Totals	*N*	%Error$_T$
Call sign recognition	18 (2.67%)	83 (12.30%)	101	1350	7.481
Error identification	90 (16.67%)	72 (13.33%)	162	1080	15.000
Prediction	60 (16.67%)	38 (10.56%)	99	720	13.750

Finally, a Wilcoxon signed-rank test was conducted to compare performance accuracy across the three tasks. The risk of making an error in the Prediction task was almost two times greater than the risk of making an error on the Recognition task (Z = –3.431, p = .001, r = .51, risk ratio = 1.84). Similarly, the risk of errors in the Error identification task was approximately twice the risk of an error in the Call sign recognition task (Z = –4.338, p < .001, r = .65, risk ratio = 1.88). There was no evidence of a statistically significant difference in performance accuracy between Error identification and Prediction tasks (p = .889). Error analysis related to the types of errors (misses and false alarms) was conducted for each of the tasks separately, and this is discussed in the following sections. Given the violated assumption of normality, a non-parametric test was used for these analyses.

8.3.3.1. Call Sign Recognition Task

To compare performance accuracy between language conditions (L2 and Mix), a Wilcoxon signed-rank test was used. The risk of making miss errors in the Mix condition was almost five times the risk of making miss errors in the L2 condition (10.67% vs. 2.22%; Z = –5.728, p < .001, r = .85, risk ratio = 4.80). There was no statistically significant difference between the Mix and the L2 conditions in making false alarms (p = .131). The miss rates were greater than the false alarm rates in both the L2 condition (2.22% vs. 0.44%; Z = –2.384, p = .017, r = .36, risk ratio = 5.00) and the Mix condition (10.67% vs. 1.63%; Z = –5.605, p < .001, r = .84, risk ratio = 6.55). Data are presented in Table 44.

Table 44

Distribution of Errors (Miss and False Alarm), Hits and Correct Rejections (CR) in Call Sign Recognition Task, Total Number of Errors across Language Conditions, and Percentage of Errors from 1350 Stimuli (%Error$_T$)

	Miss	False alarm	Hits	CR	Error Total	%Error$_T$
L2	15	3	210	447	18	
Mix	72	11	153	439	83	
Total	87	14	363	886	101	7.481

8.3.3.2. Error Identification Task

A Wilcoxon signed-rank test was used to compare performance accuracy between the language conditions on the Error identification task and indicated greater false alarm rates than miss rates in the L2 condition (9.85% vs. 3.79%; *Z* = –3.643, *p* < .001, *r* = .54, risk ratio = 2.60) and in the Mix condition (7.26% vs. 3.41%; *Z* = –2.694, *p* = .007, *r* = .40, risk ratio = 2.13). The false alarm rates were also greater in the L2 condition than the Mix condition (9.85% vs. 7.26%; *Z* = –2.247, *p* = .025, *r* = .33, risk ratio = 1.36); however, there was no evidence of a statistically significant difference between the language conditions in making miss errors (*p* = .822). Error counts are presented in Table 45.

Table 45

Distribution of Errors (Miss and False Alarm), Hits and Correct Rejections (CR) in Error Identification Task, Total Number of Errors across Language Conditions, and Percentage of Errors from 1080 Stimuli (%Error$_T$)

	Miss	False alarm	Hits	CR	Error Total	%Error$_T$
L2	25	65	155	295	90	
Mix	23	49	157	311	72	
Total	48	114	312	606	162	15.00

8.3.3.3. Prediction Task

A Wilcoxon signed-rank test revealed no statistically significant difference between miss and false alarm rates in the Mix condition (p = .802). However, in the L2 condition, the risk of missing a predicted number was almost three times greater than the risk of making a false alarm (6.37% vs. 2.52%; Z = –3.315, p = .001, r = .49, risk ratio = 2.53). Additionally, the risk of making a miss error was almost twice as high in the L2 condition compared with the Mix condition (6.37% vs. 2.67%; Z = –3.264, p = .001, r = .49, risk ratio = 2.39). The risk of making false alarm error did not differ between the L2 and Mix conditions (p = .636). Data are presented in Table 46.

Table 46

Distribution of Errors (Miss and False Alarm), Hits and Correct Rejections (CR) in Prediction Task, Total Number of Errors across Language Conditions, and Percentage of Errors from 720 Stimuli (%Error$_T$)

	Miss	False alarm	Hits	CR	Error Total	%Error$_T$
L2	43	17	137	163	60	
Mix	18	21	117	204	39	
Total	61	38	254	367	99	13.75

8.3.4. SDT Measures

To test the effect of the Background talk factor on performance in Call sign recognition, Error identification and Prediction task, the SDT measures (sensitivity, d'; criterion, C) were calculated.

Although the SNR was the same across the three tasks, findings indicated minor variations in the detectability of the stimuli in all language conditions ($1.97 \leq d' \leq 3.97$). Additionally, the findings revealed response bias towards *no* responses in the first task, suggesting that participants were more prone to missing the target number. Similar response bias toward non-detection was observed in the third task, in both the L2 (C = 0.30) and Mix conditions (C = 0.105). When identifying erroneous equations in the L2 condition, there was minor opposite bias, toward false alarms (C = –0.085). In the Mix condition for this task, the responses were almost unbiased (C = –0.02). Data are summarised in Table 47.

Table 47

Sensitivity (d') and Decision Criterion (C) across the Three Tasks (Call Sign Recognition, Error Identification, and Prediction), and Two Language Conditions, Second (English) Language (L2), and Language Switching (Mix)

		L2	Mix
	z(HR)	1.50	0.47
Task 1	*z(FAR)*	–2.47	–1.97
Call sign recognition	*d'*	3.97	2.44
	C	0.485	0.750
	z(HR)	1.07	1.14
Task 2	*z(FAR)*	–0.90	–1.10
Error identification	*d'*	1.97	2.24
	C	–0.085	–0.020
	z(HR)	0.71	1.11
Task 3	*z(FAR)*	–1.31	–1.32
Prediction	*d'*	2.02	2.43
	C	0.300	0.105

8.4. Discussion

The findings of this study indicated no effect of simultaneously listening to background talk (Chinese or English) on the speed or accuracy of call sign recognition, error identification, or prediction performance. There was no evidence of a statistically significant interaction effect of the Background talk and Language condition factors. Based on these findings, no assumptions can be made regarding the *cocktail party effect* in air traffic communications—whether pilots are able to detect their call sign when simultaneously listening to two speech sources. Based on the cocktail party effect (Neisser & Becklen, 1975), uttering an aircraft call sign—the ‘name’ of an aircraft pilots are flying—should attract pilots’ attention even when they converse with each other and the radio is not being attended to. However, based on the findings of this study, it cannot be determined whether the attention of the pilot would be attracted immediately to an unexpected radio call when they are listening to a co-pilot, or whether the language used (native or second language) can influence this effect. Yet, some Crew Resource

Management (CRM) training resources consider this effect as confirmed (Crew Resource Management, 2018).

A possible explanation for the absence of the effect of background talk can be found within the selective attention field. Participants might have assigned greater priority to the computer-based experimental tasks so that when they heard stimuli from the computer tasks, they directed attention to them. It was somewhat surprising, therefore, that they could retrieve some words, or even an entire story, from background talk they just heard. There was a large variety in responses. This contrasts with McAnally et al. (2010) and Cherry's (1953) findings but supports Demany et al.'s (2010) interpretation that detection of change is to some extent automatic.

Analysis of participants' ability to retrieve information from background talks would help to clarify how attention was switched. Unfortunately, because there were no clear criteria for analysing the answers participants provided from background talks, no statistical analysis could be performed. The reason was that the analysis of the retrieval processes was not the primary aim of this study, but to explore the effect of background talk on performance in the three tasks. Therefore, this discussion can be limited only to an assumption that participants tuned out the background talk and focused on the tasks, based on the response of the Survey Participant 76. If this assumption was correct, the question of what made it possible to retrieve some information arises. For example, in informal reflections that participants shared after they finished the experiment, one participant reported that initially, while a talk was playing, she remembered many things, but when she was about to write them down after the task finished, she forgot them. This suggests short-term memory limitations rather than factors related to attention. This would be in accordance with Endlsey (1995), who identified memory as a significant factor affecting SA. A participant who provided not only single words, but a recapitulation of the whole background talk, reported that he might have made the story up, rather than written what he really remembered. Shinn-Cunningham (2008) explains that even if all the content of one signal is not perceived, people can fill in missing snippets. Further, some of the information in a newly attended stream of speech can be missed even after a listener switches attention, because of the time it takes to switch (Shinn-Cunningham, 2008). Therefore, there might be some attention switch costs, besides language switch costs, when listening to two simultaneous speech sources, such as cockpit communication and a radio broadcast. This potentially points to other factors besides memory, such as an interference effect, which might affect participants' ability to retrieve information from background talks.

The analysis of background talk can be addressed in future research to allow more thorough analysis of the four combinations of background talk language and air traffic environment language. Until then, it remains unclear how participants prioritized their attention, or what allowed them to pick up some information from a speech source that was not attended to. The extent to which irrelevant information is tuned out by selective attention remains an open question.

Nevertheless, there was evidence of a statistically significant main effect of the Language condition factor. When recognizing a target call sign, participants performed faster and more accurately in the L2 condition than the Mix condition. However, the performance was faster in the Mix condition when identifying errors and predicting. Although the effect of language condition varied across the three tasks, what caused its effect to persist, regardless of selective attention filtering the language of a background talk, remains unclear. Answering this question would be of interest as it could better explain selective attention and SA.

The observed differences may indicate that the primary factor that affected performance was language alternation itself. However, the analysis of the switch and mixing costs does not seem to clearly support this assumption. The switch and mixing costs were only able to be analysed for the Call sign recognition and Error identification tasks, and no statistically significant switch or mixing costs were found in the Call sign recognition task. In the Error identification task, only mixing costs were observed and they were symmetrical; that is, it was faster to switch to a Chinese stimulus after hearing a stimulus in English language than the other way around. This is contrary to the findings of previous research (Bobb & Wodniecka, 2013; Costa & Santesteban, 2004a; Meuter & Allport, 1999). Information about participants' English language proficiency could provide necessary information for the explanation of these findings, yet, because the data was incomplete, the analysis was not performed. Therefore, whether the primary effect can be explained by language alternation rather than attention change remains an open question for future research.

When recognizing a call sign (level 1 SA), a bias toward *no* responses (miss errors) was observed, which was greater in the Mix condition than in the L2 condition. Similar bias was found in the Prediction task (level 3 SA), where slightly more errors were made in the L2 condition. Specifically, missing the predicted number occurred slightly more frequently in the L2 condition than the Mix condition, yet false alarm errors did not differ between the conditions. When identifying an error (level 2 SA), the bias was more toward *yes* responses (false alarms). However, no difference in accuracy was found between the language conditions. As before (see section 6.4), different bias in the Error identification task compared with the Call sign recognition and Prediction tasks, could be attributed to the nature of the instruction, which was to identify an error. However, there appears to be no sufficient information in this study to explain the opposite finding of more accurate performance in the L2 than in the Mix condition across the three levels of SA. More research would be needed to explore the practical implications for aviation safety. Owing to the primary focus of this study on the effect of background talk—which did not appear to affect performance accuracy—no further detailed discussion is provided.

There were four potential limitations of the current study. The first relates to the development of the experimental tasks and methodology. The remaining three limitations are the same as those in Studies 2, 3 and 4, as the present study used material largely from those sources.

The first limitation involves the methodology of the experimental tasks. One concern is the use of only numeric call signs and whether similar results would be obtained for alphanumeric call signs. Next, numerical equations may not precisely simulate the real-life errors of air traffic communication messages. Nevertheless, this study presents a simplified analogy of problem solving involving numbers. Additionally, during development of the stimuli for the Mix condition of the Prediction task in Study 5, a mistake was made whereby three, instead of four, predicted numbers were present (*yes* response) in a sequence. This created an imbalanced stimuli list; that is, the proportion of present and absent predicted numbers in the L2 condition was equal (4 were present and 4 absent), whereas in the Mix condition, the predicted number was present only infrequently (3 present and 5 absent). Also, the number sequences might not have precisely simulated the prediction process involved in obtaining SA. The number sequences are simplified approximations of the cognitive processes involved in prediction. More importantly, three different intervals between the numbers within a sequence could affect the accuracy of the RT measurement. These limitations would need to be addressed in future research. Caution is therefore necessary when attempting to generalize the findings.

The second limitation, the population, specifically, the inter-individual variability between non-aviation students, the participants, and the aviation personnel, might also influence the generalization of findings. However, as discussed before, the problems raised by this possibility are not unique to aviation psychology studies (Barshi & Farris, 2003).

The third limitation relates to the self-reported IELTS test scores. For several reasons, the self-reported IELTS test scores did not reflect the English language literacy of the participants and were therefore excluded from the data analysis. For example, some of the IELTS tests had been completed 10 years previously, and thus the scores were no longer valid (British Council, n.d.e). Some participants did not complete the IELTS, because they were doing an internship programme and there was no requirement for them to be tested. In these cases, the data were completely missing.

The insufficient data of participants' English language proficiency provided a lesson. To obtain a valid and reliable understanding of participants' English language proficiency, an English test should have been conducted for every participant. However, a less time-consuming test could be chosen or developed, given that in this study, the IELTS test would have lasted longer than the experiment itself, even though English proficiency was not the primary focus. In addition, it would be desirable to choose a test that could also be performed by the aviation personnel population to allow English language proficiency alignment. This would allow the development of a standardized test that could be used for research purposes in language-focused aviation psychology studies as a simplified alternative to the standard assessment, meeting the requirements for the ICAO Rating Scale.

The fourth limitation relates to several concerns about the background talks. Participants were asked to write everything they could remember from a talk in English, regardless of whether the talk was in English or Chinese. Some information could, therefore, have been lost or distorted between participants translating what they had heard in Chinese and writing their response in English. Participants were allowed to use a dictionary to translate Chinese words into English. However, very few participants chose to do so.

The background talk and experimental tasks were played from two different computers, which may not have been of equal volume. This could have affected the placement of the decision criterion (*C*), in terms of the speech stream they assigned greater priority to, specifically, whether they put more emphasis on the computer task and paid less attention to the background talk. However, in the instructions, background talk and the experimental tasks were stressed as equally important, and this limitation should not have caused any adverse effects on performance, also because in aviation, the speech streams always come from different sources and locations. For details on the effect of the location of speech sounds see Drullman and Bronkhorst (2000), who found that spatial separation had a positive effect on communication, or Shinn-Cunningham (2008), who summarised that the more distinct competing streams of speech are from one another, the more complete the suppression of the stream in the perceptual background. In other words, participants were less likely to confuse words across streams, but also recall fewer words (Shinn-Cunningham, 2008). The use of a sound level meter could correct this limitation in future studies.

Reflecting upon the discussion and limitations, it is reasonable to expect that some effect of background talk in real life can be observed. For further research in this area, attention could be directed to the data presented in Table 36, despite no evidence of a statistical significant interaction effect of the Background talk and Language condition factors being observed. This data may indicate where the possible presence of the interaction effect might be detected, when more thorough research, corrected for the identified limitations, has been performed. The fastest responses might be observed in monolingual situations, where background talk and the air traffic environment use the same language, and slowest in bilingual air traffic environments combined with the native language of a conversation of a crew during the flight (see Table 36). Should further research confirm this speculation, the practical implications for practice and safety could be crucial, given that the combination of bilingual air traffic and native-language cockpit talk is likely the most common experience. However, because of the limitations of this study, the discussion will not go into further detail and no clear implications for aviation practice can be provided. As such, the guiding research question for this study was only partially answered.

In conclusion, the findings of the current study suggest that background talk did not have any effect on Call sign recognition (level 1 SA), Error identification (level 2 SA), or

Prediction (level 3 SA) performance in this study group. However, there was a difference in speed and accuracy of performance between the language conditions, which varied across the three levels of SA.

CHAPTER NINE

Study 6: Sterile Cockpit

9.1. Introduction

The series of studies in this thesis was designed to explore the research question *whether bilingual language perception adversely affect performance of bilinguals*. Attention was concentrated upon the effect of monolingual versus bilingual language conditions on three cognitive tasks that underlie the three levels of SA, namely, recognition, comprehension and prediction. These were investigated in studies 2, 3 and 4. Study 5 sought to investigate the additional effect of background talk analogous to a cockpit conversation between pilots on performance. In Study 1, it was found that pilots and ATCOs communicate with their colleagues in their native language while on duty, switching their attention back to radio broadcasting when necessary. Therefore, background talk has the potential to distract pilots' attention and impair their task performance. After analysis of a series of accidents where causes were attributed to the distraction of the flight crews from their flying duties, in 1981 the FAA introduced the formal requirement for a sterile cockpit (Baron, 1995), which addresses the background talk factor.

The sterile cockpit rule reads: *"No certificate holder shall require, nor perform any duties during a critical phase of flight except those duties required for the safe operation of the aircraft… any activity which could distract any flight crewmember from the performance of his or her duties or which could interfere in any way with the proper conduct of those duties… engaging in non-essential conversations within the cockpit and non-essential communications between the cabin and cockpit crews, and reading publications not related to the proper conduct of the flight are not required for the safe operation of the aircraft"* (FAA, 1981; 14 CFR 121.542 - Flight Crew Member Duties, p. 202), Less precise is the European Regulation (EASA) No 965/2012 description of "*any period of time when the flight crew members shall not be disturbed…* (in order) *to increase the flight crew members' attention*".

Typically, the sterile cockpit rule prevents the flight crew from engaging in unnecessary talk (analogous to the background talk in Study 5 of this thesis) below an altitude of 10,000 feet (FAA, 1981); that is, the rule is followed during take-off and landing, and is not required during the cruise phase of flight. A wide spectrum of activities can be considered non-adherence to the sterile cockpit rule. However, the communication aspect, in particular "*engaging in non-essential conversations within the cockpit and non-essential communications between the cabin and cockpit crews*" is relevant to the current study.

Sumwalt's (1993) analysis of ASRS reports revealed that the most cited non-adherences to the sterile cockpit rule were related to conversations rather than performance of some kind of a task. They were: (i) extraneous conversation between cockpit crew members; (ii) the captain of an aircraft admitting to conversation not pertinent to flying duties; (iii) extraneous conversation with a person sitting on a jump seat; (iv) extraneous conversation with ATCOs; (v) extraneous conversation with flight attendants; (vi) non-pertinent company radio calls, and, (vii) passenger announcements. According to the Sumwalt's (1993) review, the first and main activity that decreased performance during times that required non-disturbance was cockpit communication (termed background talk in this thesis). In this context, the contribution of this comparative-contrasting study can be crucial.

To investigate the importance of the *sterile cockpit* for bilinguals in a bilingual air traffic environment, the current study compared the findings of Studies 2, 3 and 4 with those of Study 5. Even though no evidence of an effect of the Background talk factor on performance was found in Study 5, there was some evidence that the language condition in which participants performed the tasks affected the speed and accuracy of responses. This study sought to explore whether these effects were different from the effects found in Studies 2, 3 and 4, where the same tasks were performed without background talk. To do so, the following guiding research question for Study 6 was formulated:

Question 6: Is there a difference in performance speed and accuracy between the tasks performed with and without background talk?

Exploring differences between the studies can point to the areas that background talk affects, which may help pilots and ATCOs to perform their tasks with greater awareness and accuracy.

9.2. Method

9.2.1. Overview

Because the present study was a comparative analysis of the previous experimental research, the Method section focuses only on the development of the data analysis. To explore the effect of the presence of background talk on performance of recognition, comprehension and prediction—the three cognitive processes involved in situation awareness—data from the experiments conducted without background talk were contrasted with the data from the experiment in which these three tasks were conducted while participants simultaneously listened to background talk, either in their native language (Chinese), or second language (English). Prior to the analysis, the method of each of the experiments was thoroughly reviewed to decide whether the comparison was methodologically feasible. This is discussed in the following section. Subsequently, the choice of a suitable design and analysis was made.

9.2.2. Considerations for the Analysis

To assess the potential validity of the proposed comparison, three factors were considered. First, the number of stimuli differed between the experiments. The Call sign recognition task used 5100 stimuli without background talk, and 1230 stimuli with background talk. The Error identification task used 5200 stimuli without background talk and only 984 with background talk. The Prediction task used 1920 stimuli without background talk, and 656 stimuli with background talk. The difference between the number of stimuli in tasks with and without background talk could affect the statistical power of this comparative study; that is, the probability of concluding there is no effect when an effect exists (type II error).

Second, the native (Chinese) language condition (L1) was not used when the tasks were performed with background talk. Therefore, the L1 condition was not included in this comparison analysis.

Third, in Study 4 (Prediction without background talk), the Position factor had 4 levels and the Pattern factor had 2 levels, whereas in the Prediction tasks of Study 5 (with background talk), the Position factor had only 2 levels, and used only a simple pattern for number sequence development. Both Position and Pattern factors were found to be statistically significant. Consequently, the mean pure RTs obtained in Study 4 were generally longer than the mean pure RTs measured in Study 5. To compare the findings of Study 4 and the third task of Study 5, therefore, the mean pure RT of Study 4had to be adjusted. This was done by computing an adjusted mean pure RT that considered only Positions 1 and 2 and Pattern 1. The comparison of findings of the two studies (Study 4 and 5) was then considered as reasonable, given that these potential limitations are identified for readers.

9.2.3. Measures

The dependent variables were the RT and the number and type of errors (miss and false alarms), which were compared between studies without background talk (Studies 2, 3, and 4) and their corresponding tasks with background talk (Study 5). Thus, the Background talk factor was a between-subjects independent variable with three levels: background talk in Chinese, background talk in English, and no background talk. Each of the levels of the between-subjects variable was compared across the within-subjects independent variable, the Language condition factor. The Language condition factor had two levels: the monolingual English language (L2) and the language switching (Mix) condition. The comparison was made for each of the three tasks separately; that is, Call sign recognition, Error identification and Prediction.

Consideration regarding which statistical test is most suitable for the data analysis is necessary. Taking into account the measures and the research question used in this study, the relative merits of a mixed ANOVA and an independent-samples *t*-test are discussed. The mixed ANOVA is an omnibus test with the primary purpose to identify an interaction between a within-subjects factor (Language condition factor with 2 levels: L2 or Mix) and a between-subjects factor (Background talk factor with three levels: Chinese background talk, English background talk, and No background talk) on the dependent variable (RT). However, the interaction effect between these factors is not the primary aim of this study. Moreover, provided that the between-subjects factor has three levels, the mixed ANOVA cannot specify which of these three levels are significantly different from each other. Determining where the differences lie is the primary aim of this study. For this purpose, an independent-samples *t*-test might be more suitable as it compares the means of two independent groups in order to determine whether they are statistically significantly different. In other words, it allows an analysis of simple main effects, instead of the interaction effect. Therefore, an independent-samples *t*-test was used for the data analyses.

9.3. Results

9.3.1. Overview

The Results section is organised in line with the two main indices of performance; speed and accuracy. Comparison of the SDT measures is provided at the end of the results section.

9.3.2. Performance Speed

To compare performance speed in the three tasks (Call sign recognition, Error identification and Prediction) when listening to background talk and without background talk (analogous to a sterile cockpit), data from Studies 2, 3 and 4 were compared with data from Study 5. Data for both performance speed and accuracy with and without background talk across the three tasks and two language conditions are summarised in Table 48.

Table 48

Mean Pure Response Times in Seconds (M), Standard Deviations (SD), Error Counts and Percentage of Errors within Conditions (%Error) of Call Sign Recognition (Task 1), Error Identification (Task 2), and Prediction (Task 3) Tasks with and without Background Talk

		Task 1		Task 2		Task 3	
		L2	Mix	L2	Mix	L2	Mix
Without background talk	*M*	0.228	0.339	1.563	1.469	7.418	4.272
	SD	0.256	0.254	0.738	0.588	2.748	2.636
	Errors	17	14	123	122	74	53
		(1700)	(1700)	(1600)	(2000)	(640)	(640)
	%Error	1.00	0.82	7.69	6.10	11.56	8.28
With background talk in Chinese	*M*	0.632	0.798	1.828	1.417	6.114	5.749
	SD	0.486	0.540	0.861	0.663	4.42	3.895
	Errors	9	44	46	44	32	18
		(345)	(345)	(276)	(276)	(184)	(184)
	%Error	2.61	12.75	16.67	15.94	17.39	9.78
With background talk in English	*M*	0.538	0.711	1.559	1.178	5.149	4.559
	SD	0.520	0.461	0.728	0.394	3.976	3.506
	Errors	9	39	44	28	28	21
		(330)	(330)	(264)	(264)	(176)	(176)
	%Error	2.73	11.82	16.67	10.61	15.91	11.93

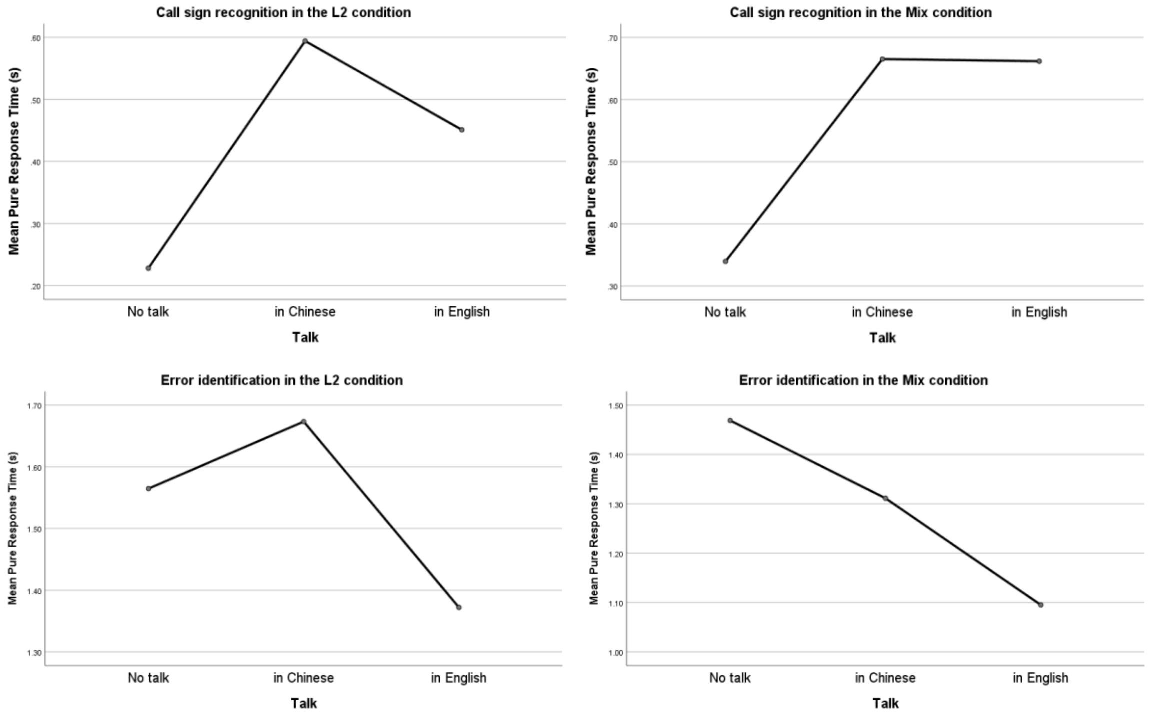
Note. Total number of stimuli within language conditions for each task is presented in parentheses.

Prior to testing whether the differences in performance speed with and without background talk were statistically significant, the assumption of normality was evaluated and was found to be satisfied as the distributions were associated with skewness and kurtosis within the span of –2.0 and +2.0 (Cramer, 1998; George & Mallery, 2010). Owing to the satisfied assumptions, an independent-samples *t*-test was conducted to compare performance speed with and without background talk across the language conditions.

In the Call sign recognition task, the performance was found to be statistically significantly faster without background talk in both language conditions (L2 and Mix), regardless of the language of the background talk (Chinese or English). In the L2 condition, the performance was faster without background talk (228 ms) than with background talk in both Chinese (632 ms; *t*(52) = –3.616, *p* = .001, *r* = .43), and English (538 ms; *t*(51) = –2.284, *p* = .027, *r* = .29). Similarly, in the Mix condition, recognition of call signs without background talk was faster (339 ms) than with background talk in both Chinese (798 ms; *t*(51) = –3.931, *p* < .001, *r* = .47), and English (711 ms; *t*(50) = –3.290, *p* = .002, *r* = .40).

In the Error identification task, performance was found to be statistically significantly faster in the Mix condition with background talk in English (1.178 s) than without background talk (1.469 s; *t*(57) = 2.594, *p* = .012, *r* = .37). No other comparisons were found to be statistically significant (.589 ≥ *p* ≥ .312).

In the Prediction task, no statistically significant differences were found between performance with and without background talk in the Mix condition regardless of the language of the background talk (with Chinese talk, *p* = .084; with English talk, *p* = .536). In the L2 condition, performance was faster with the background talk in English (5.149 s) than without background talk (7.418 s; *t*(58) = 2.331, *p* = .023, *r* = .29), and faster with the Chinese background talk (6.114 s) than without it (7.418 s; *t*(58) = 3.552, *p* = .001, *r* = .43). Profile plots are presented in Figure 16.


Call sign recognition in the L2 condition

Call sign recognition in the Mix condition

Error identification in the L2 condition

Error identification in the Mix condition

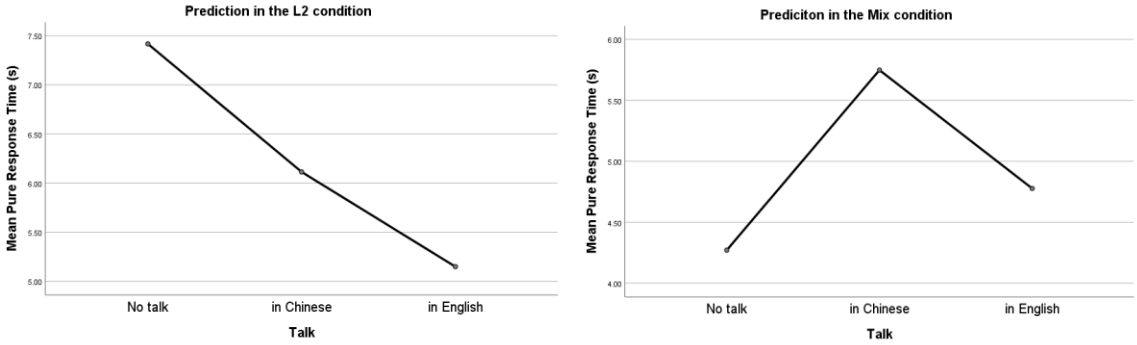


Figure 12. Profile plots for Background talk factor across the three tasks (Call sign recognition, Error identification, Prediction), and two language conditions (L2, Mix).

9.3.3. Performance Accuracy

Prior to testing whether performance accuracy differed between the tasks with and without background talks, the assumption of normality was evaluated and was found to be violated as the distributions were associated with skewness and kurtosis beyond the span of –2.0 and +2.0 (Cramer, 1998; George & Mallery, 2010). Therefore, a non-parametric test was used.

In the Call sign recognition task, Mann-Whitney U tests revealed no statistically significant difference between the groups with and without background talk in the L2 condition, nor with background talk in Chinese (*p* = .319), or in English (*p* = .776). The same task in the Mix condition was statistically significantly more accurate without background talk in Chinese (0.82% vs 12.75%; *U* = 23.00, *p* < .001, *r* = .83, risk ratio = 13.848), as well as in English (0.82% vs 11.82%; *U* = 69.00, *p* < .001, *r* = .72, risk ratio = 12.940). However, no statistically significant difference was found between the effects of the two languages of background talk, Chinese and English (*p* = .876).

In the Error identification task performed in the L2 condition, the performance was significantly more accurate without background talk in Chinese (7.69% vs 16.67%; *U* = 293.500, *p* = .041, *r* = .26, risk ratio = 2.001), and English (7.69% vs 16.67%; *U* = 246.00, *p* = .006, *r* = .35, risk ratio = 2.001). Likewise, in the Mix condition, the performance was found to be significantly more accurate without background talk in Chinese (6.10% vs 15.94%; *U* = 161.00, *p* < .001, *r* = .54, risk ratio = 2.392), and English (6.10% vs 10.61%;

U = 287.00, *p* = .023, *r* = .29, risk ratio = 1.668). However, no significant difference was found between the effects of the two languages of background talk, Chinese and English (*p* = .069).

No statistically significant difference in accuracy was found between performing the Prediction task in the L2 condition without background talk and with background talk in Chinese (*p* = .360), or English (*p* = .272). Likewise, in the Mix condition, the analyses revealed no statistically significant difference in accuracy between the groups (with Chinese talk, *p* = .629; with English talk, *p* = .622).

9.3.4. SDT Measures

To compare the internal response and effect of noise with and without background talk, the SDT measures (sensitivity, *d'*; criterion, *C*) of Studies 2, 3, and 4 were compared with the SDT measures of Study 5. Data are presented in Table 49.

Table 49

Comparison of Sensitivity (D') and Decision Criterion (C) of the Three Tasks Conducted with and without Background Talk in Second (English) Language (L2), and Language Switching (Mix) Conditions

		Task 1		Task 2		Task 3	
		L2	Mix	L2	Mix	L2	Mix
Without background talk	*d'*	4.75	4.69	3.07	3.26	2.39	2.84
	C	0.425	0.215	–0.245	–0.190	0.045	0.190
With background talk	*d'*	3.97	2.44	1.97	2.24	2.02	2.43
	C	0.485	0.750	–0.085	–0.020	0.300	0.105

Based on the comparisons of sensitivity (*d'*) between tasks with and without background talk, the discriminability of stimuli from background noise had the tendency to decrease in all tasks with background talk by an average of 0.988, despite the noise in these tasks being set at the same level as in the Prediction task without background talk (Study 4), and even lower compared with the Error identification task without background talk (Study 3).

The findings also indicated a shift in *Criterion* (response bias) when listening to background talk. Specifically, when listening to background talk in the Call sign recognition task, participants' tendency to miss (*no* response) a target call sign increased in both the L2 (*C* increased from 0.425 to 0.485) and the Mix (*C* increased from 0.215 to 0.750) conditions. The tendency to miss a number in the Prediction task also increased in the L2 condition (*C* increased from 0.045 to 0.300) when simultaneously listening to background talk but decreased in the Mix condition (*C* decreased from 0.190 to 0.105). In the Error detection task, the responses became less biased in both the L2 (*C* decreased from -–0.245 to –0.085) and Mix conditions (*C* decreased from –0.190 to –0.020) when background talk was present. The largest difference in the shift of response bias was in the Call sign recognition task in the Mix condition, where the tendency to miss a target greatly increased (by 0.535) when listening to background talk.

9.4. Discussion

In this study, the idea of a sterile cockpit was used to provide context—that is, as an analogy of a situation where there was no background talk (as opposed to when background talk was invoked, which would be analogous to not adhering to the sterile cockpit rule)—to explore its effect on call sign recognition, error identification, and prediction performance. The background talk condition was considered to represent the kinds of communication that might occur between two colleagues during a non-critical phase of flight (for example, talking about work issues not related to flying, such as rostering).

Nevile (2004) stressed that despite communication having been studied since the beginning of aviation research, it has focused on communication in non-routine situations, accidents and incidents (e.g., Cushing, 1994), or focused on the type and construction of utterances (e.g., Barshi & Farris, 2013). Yet, how pilots and ATCOs "routinely communicate in their ongoing interactions with one another as they perform the typical tasks" (Nevile, 2004, p. 12) has not been explored. This study sought to add to this relatively sparse research area, by focusing on what happens in cognition when pilots listen to a general conversation in routine situations during long flights and an unexpected radio communication occurs. The sterile cockpit rule was chosen to provide the context, despite it applying only below an altitude of 10,000 feet (FAA, 1981); that is, not *en* route in the cruise phase.

When recognizing a target call sign, performance was found to be faster and more accurate when participants were not simultaneously listening to background talk. The presence of background talk also shifted participants' response bias toward *no* responses; that is, the misses of a target call sign increased. This response bias shift was found to be greater in the Mix condition than the L2 condition. In fact, this was the largest adverse effect of the background talk on participants' response bias observed across the tasks and language conditions. The findings are not surprising when considering them within the first level of situation awareness (level 1 SA). Specifically, informal talk can distract pilots, so they more likely to miss an ATC message. Non-adherence to the sterile cockpit rule might be evident in the 1988 crash of the Dash 7 into high terrain in Norway, known as the Torghatten Accident (Accident Investigation Board Norway, 2013). Among the other contributing factors, the pilot was having a conversation with a passenger on the jump seat (Flight Safety Digest, 1992).

In the Error identification task, the findings were somewhat ambiguous. In the L2 condition, no difference in performance speed or accuracy was found when performing the task with or without background talk. In the Mix condition, the performance was more accurate without background talk, but, surprisingly, faster with background talk in English. No difference was found when the talk was in Chinese. This may suggest that their native Chinese language distracted the participants from identifying errors more than their second language, English. Presumably, because of its dominance, even when the background talk was not attended to, it cannot be completely filtered out. This explanation is in accordance with the findings reported by Hodgetts at al. (2005) that meaningful information increased pilot workload and impaired flight task performance. Hypothetically, if pilots were talking with each other in their native language while operating in bilingual air traffic environment, their conversation would not affect their reaction speed, but the accuracy of these reactions would be impaired.

Somewhat surprisingly, simultaneously listening to background talk had a positive effect on response bias in the Error identification task—it decreased, and the responses became almost unbiased. Decrease in bias, however, does not necessarily mean a decrease in the overall number of errors, but rather that the occurrence of miss and false alarm types of errors have become similar and either error type may dominate. Together, the findings of the Error identification task may imply that the comprehension of communication (level 2

SA) is better in the monolingual English air traffic environment. Also, because of the observed differences proposed in the Mix condition above, the importance of the sterile cockpit can be stressed in the bilingual air traffic environment.

The effect of background talk in the L2 and the Mix conditions of the Prediction task were the opposite of those in the Error identification and Call sign recognition tasks. Performance accuracy seemed to be unaffected by simultaneously listening to background talk, compared with the sterile cockpit (tasks without background talk), in both the language conditions. The speed of predicting in the bilingual air traffic environment did not differ either. Somewhat counter-intuitively, however, the speed of making predictions was found to be faster when simultaneously listening to background talk, in both language conditions. It can be speculated that participants might tune out the background talk completely, and focus on prediction only, because to predict they had to extrapolate the pattern that the number sequences followed, which required simple mental calculations. Moreover, the presence of background talk increased the response bias in the L2 condition but decreased it in the Mix condition. A potential explanation for the effect of background talk on performance can be found in a real-life situation, obtained from ASRS report 167026. The report serves as an example of both Error identification and Prediction tasks, given that the pilot recognized that something was wrong, which was the potential future hazard. The report reads as follows:

"*While descending into a broken deck of clouds, unannounced traffic appeared at 12 o'clock and less than a mile, climbing up our descent path. In my best estimation we were on a collision course. I immediately, without hesitating, instinctively pushed the aircraft nose down and to the right to avoid impact. The captain was engaged in a conversation with [somebody] on the jump seat.*" (Sumwalt, 1993, p. 19).

The findings of this study could imply that level 3 SA in the bilingual air traffic environment is unaffected by non-adherence to the sterile cockpit (predicting in the presence of background talk) and can positively decrease the response bias. However, ASRS report 67026 (Sumwalt, 1993), potentially suggests a distinguishing feature. The pilot who noticed the aircraft on a collision course, did so when background talk was present but he was not involved in the conversation—the captain was talking with someone on the jump seat. However, the captain, who was actively involved in a conversation, did not notice the aircraft. This may indicate a difference between the potential effects of listening only, and listening and speaking, on the allocation of attention, and thus, also, on task performance, and SA. In the present study, participants were only listening to two simultaneous speech stimuli—they did not talk themselves. Therefore, inferences for aviation based on the findings of this study, should be approached with caution, primarily because neither prediction nor error identification can occur without recognition, which was found to be faster and more accurate without background talk.

The finding that the difference in performance appeared only when background talk was present, but with no evidence of a difference between the two languages of background talk, may be understood by referring to language switching theory. Specifically, language switching causes additional processing time, which explains the performance speed difference (e.g., Declerck et al., 2012) and is attributed to having to overcome the residual inhibition of the non-used language (Green, 1998). In this study, performance was found to be faster and more accurate without background talk regardless of whether it was in the native (Chinese) or second (English) language. In other words, even when both task and background talk were in the same language, and, hence, participants were not required to switch between the languages, performance with background talk was generally found to be worse than without background talk. Therefore, it can be assumed that it was not the alternation between different languages—the language of background talk and the language of the task—that caused the difference in performance, but rather the presence or absence of background talk. This suggests that it was the task switching (from background talk to a task and then back to the talk again) that caused performance to decrease. This would be in accordance with Weissberger et al. (2015), who compared the neural correlates of task switching and language switching using functional magnetic resonance imaging (fMRI) analysis, and found that bilinguals have greater efficiency for sustaining the inhibition of the non-target language than the non-target task (see also Branzi et al., 2016; Declerck et al., 2017; Liu et al., 2016).

Dual language processing and language switching from Chinese background talk to English task does not seem to be the major factor affecting the performance of bilinguals when listening to two simultaneous speech stimuli. Rather, it is abstaining from background talk, and thus, applying the sterile cockpit rule, which may benefit performance in both monolingual English and the bilingual air traffic environment. Because the difference in performance can be attributed to selectiveness of attention and distraction (e.g., Mattys et al., 2012) caused by the presence of background talk, the results could also be generalized to ATC, implying a need for a similar sterile control room rule. For example, anecdotal evidence suggests that ATCOs may talk between themselves about a variety of issues unrelated to their work during quiet periods.

The findings raise the question of why no evidence of a statistically significant main effect of Background talk factor was found in Study 5. Of course, the use of inferential statistics, even with power of .80, means there is a 20% chance of type II error. The rationale proposed above—that it might be the task switching rather than language switching that affected the performance—can explain it in a sense; in Study 5, there was small number of stimuli and, thus, insufficient task switching to detect a difference or effect, should it exist. Only in comparison to the previous experiments without background talk could a difference be observed. In addition, the results of this comparative study slightly varied. There was also a lack of literature addressing this topic, preventing any further comparison. Consequently, no firm conclusions can be made about the effect of the language of background talk on SA in monolingual and bilingual air traffic environments. This leads to consideration of the potential limitations of this study.

The limitations of this comparison-contrasting study lie in the methodological differences between the studies, which were described in the method section. Namely, in the first task, Call sign recognition, the Similarity factor was not used to develop the stimuli used in Study 5. Next, two factors from Study 4, Prediction, were not used in Study 5 either. In Study 5, the Pattern factor had only one level, while Study 4 had two, and the Position factor had only two levels, instead of four used in Study 4. Even though some adjustment of the RT was made to Study 5 to address this, Studies 4 and 5 cannot be considered completely equal. The same is true for Studies 2 and 3 in comparison to Study 5. Nevertheless, these limitations can be addressed in future research.

In conclusion, the findings of this study provided empirical support for the need for the sterile cockpit rule, especially for maintaining level 1 SA in bilingual language environments. Further research is required, and may also be applied to an ATC environment.

CHAPTER TEN

General Discussion

10.1. Summary

The studies comprising this thesis sought to examine the effects of bilinguals' language switching on their performance in three tasks, to investigate whether different language conditions can facilitate faster and more accurate responses, and eventually determine better SA for bilinguals. Bilingual air traffic environments are highly common in aviation, so the topic of this thesis is both relevant and important. The first study utilised an online survey as a means of exploring current experiences related to the use of language in aviation, and focused on bilinguals' language alternation and its effects on their SA. The second study employed a computer-based experiment to investigate call sign recognition in bilingual versus monolingual conditions, which represented the cognitive mechanism involved in the first level of SA. The third study continued the computer-based experimental methodology to explore the second level of SA, the ability to identify an error in a message, which is based on comprehension. The fourth study analysed the third level of SA, prediction, using a number series test. The fifth study concluded the analyses by performing these three tasks in a situation in which participants simultaneously listened to two sources of speech, performing the tasks with and without background talk, a situation considered analogous to in-flight cockpit communication, or control room communication. Finally, Study 6 compared the findings from Studies 2, 3 and 4 without background talk, with Study 5, with background talk. This chapter summarises the key observations in relation to the wider literature and considers the potential implications for the aviation industry. The discussion, therefore, addresses the central research question of this thesis:

Are there differences in the performance of bilinguals when communicating in one language and when switching between two languages?

10.2. Discussion

To investigate performance differences between monolingual and bilingual air traffic environments both quantitative and qualitative methodology were utilised. Study 1 sought to generate ideas for the quantitative research methodology. All four experiments in this thesis used the same general paradigm. Using the language switching paradigm in yes–no acoustic speech stimuli detection tasks, the three levels of SA of Chinese–English bilinguals were explored from a cognitive perspective. A fruitful way to discuss the data is to assess what the studies of this thesis contributed to the central research question.

The aim of Study 1 was to explore the current situation in aviation in relation to bilingualism, and to consider what has changed since the implementation of the LPRs. Besides the use of two languages for radio communications, Study 1 also explored the use of languages within cockpit- or control room-communications. Approximately 86% of non-native English speaking pilots and ATCOs who participated in this study achieved the higher levels 5 (46.2%) and 6 (40%) of the ICAO Rating Scale, indicating Extended and Expert English proficiency. In 2009, Prinzo and Thompson found that approximately 94% of the pilots achieved an overall language proficiency rating (LPR) of 5 (Extended), and the remaining pilots' LPR was 4 (Operational). It can therefore be assumed that the English language proficiency of aviation personnel has likely improved, and the effect can be attributed to ICAO's initiative (ICAO, 2010).

Although Study 1 revealed that monolingual English radio communications were most commonly experienced, operations in a bilingual air traffic environment, when it occurred, impaired the SA of both English-only speakers and bilinguals. The main problem that bilinguals reported in relation to bilingual air traffic was code switching (i.e., using a word in one language when communicating primarily in another), which led to the need to repeat transmissions. It was unclear whether the use of native language for cockpit or control room communications could also contribute to the reported code switching, or was merely caused by the confusion of languages for radio communication. Code-switching was found to be the cause of the language issue in Tenerife accident (Tajima, 2004), where the captain used the grammar of his native language while speaking in English. This example suggests that the problems with code-switching can be more serious than the pilots and ATCOs participating in Study 1 realized. Participants reported only obvious use of words spoken in another language and might have not considered the use of incorrect grammar. Decreased SA and the out-of-the-loop phenomenon reported by English-only participants were in accordance with the previous findings by Prinzo et al. (2008, 2010a, 2010b, 2011). Together, these raise a question about the ways in which bilingual aviation communications can be beneficial for bilinguals when their proficiency in English has improved and the use of bilingualism for radio communication represents a minor experience. This question was addressed in the subsequent empirical analyses.

Study 2 provided evidence that the recognition performance was faster in the monolingual English language condition than the bilingual condition, but no difference in performance accuracy between the two conditions was found. Performance was largely affected by the similarity of the numeric call signs, with increased similarity causing longer latencies and more errors. The findings of asymmetrical mixing costs were in accordance with those of previous studies (e.g., Campbell, 2005; Meuter & Allport, 1999; Green, 1998; Los, 1996; Verhoef, Roelofs, & Chwilla, 2009), suggesting participants had to overcome greater residual inhibition for the more dominant native language (Meuter & Allport, 1999). These findings provided support for the monolingual English language air traffic environment.

Study 3 provided contrasting findings to those of Study 2. No differences in performance speed or accuracy were found between the monolingual English and bilingual conditions, but the fastest and most accurate performance was found in the L1 condition. This finding was unsurprising, given that the task tested comprehension. Comprehension can be expected to be the easiest in an individual's native language. Avery and Ehrlich (1992, p. xv) suggested that non-native English speakers' native language affects their ability to hear English words: "the word is heard through the sound system of the native language... Sounds which occur in the native language will be heard rather than the actual sounds of English." This finding is in accordance with previous research (e.g., Campbell, 2005; Miller et al., 1995, 2000; Pavlenko, 2014) and could also be attributed to the way Chinese and English numbers are generated (Ifrah, 2000).

Importantly, performance was found to be slower when identifying an error in a message than when responding to a correct message. This finding might add to the discussion of the potential reasons for misdetection of erroneous read-backs proposed by Prinzo et al. (2006), discussed in more detail in section 6.4. The detection of infrequent erroneous read-backs would present additional cognitive processing effort and cost. Therefore, because of the large number of transmissions, ATCOs' cognitive resources may be reserved for issuing commands, rather than controlling the read-backs for potential errors. McMillan (1998) stressed that the demands on ATCOs while communicating are very high. It might be necessary to invoke a specific intention to focus on error identification to spot and correct errors. There is insufficient information available from the data to suggest that error detection would be an automatic process. In fact, it was hypothesised that the character of the instruction likely affected the response bias. Even in routine ATC communications, ATCOs filter the large number of radio communications, because delivering instructions and listening to pilots' read-backs is time demanding (McMillan, 1998). To mitigate the risk related to read-back errors, effort was put into investigating the feasibility of a read-back error detection capability that uses automatic speech recognition technology (Chen et al., 2017).

Study 3 also indicated differences that were attributed to the participants' English language proficiency. It was found that participants with English proficiency lower than IELTS Listening score 6.0–6.5, made more errors. In section 5.2.4.2.1, which discussed the IELTS and ICAO Rating Scale alignment approach, IELTS level 6.0 was referred to as roughly comparable to the minimum required Operational level 4 of the ICAO Rating Scale (Cambridge English, 2016; Harcourt Assessment, Inc., 2006). Therefore, this finding can support the alignment approach. However, this study was the only one of the four experiments in which the effect of English language proficiency was observed. Consequently, no assumptions about the effect of English language proficiency on performance in monolingual versus bilingual conditions have been made.

The occurrence of a request to repeat a stimulus was found to be four times greater in the L2 and the Mix conditions than the L1 condition, with no difference found between the L2 and Mix conditions. The finding suggests that when individuals communicate in their native language, they require fewer message repeats. In bilingual air traffic environments, repeated messages have the potential to cause increased workload, and to reduce SA. If non-native English speakers were able to communicate in their native language, the number of requests to *Say again* would potentially be lower. Therefore, a request to repeat a transmission is not necessarily a sign of a lack of professional flying or communicating skills. However, the finding also suggests that bilingual conditions, where bilinguals can use their native language as well as English, are not particularly beneficial for non-native English speakers in terms of number of requests to repeat an ATC message.

Study 4 revealed no differences in speed or accuracy of prediction performance between the language conditions. Performance was not faster or more accurate in the L1 condition. The only effects found were those of the Pattern and Position factors. When predicting, the task and its complicacy increase the demands on cognitive processing, with predictions for complicated tasks that were further ahead taking longer and becoming less accurate. These findings are in accordance with previous research (Banbury et al., 2004). One of the key contributions this study made was to provide strong evidence of the effect of Position factor on performance accuracy; beyond the cut-off for Position 2, prediction performance accuracy started to decrease. In practice, accuracy appears to be limited to predictions up to only two steps ahead. This study also provided a finding that can be interpreted in favour of the monolingual English air traffic environment; the response bias was found to be smaller in the L2 condition (0.045) than in the L1 (0.155) and the Mix (0.190) conditions. However, there is a lack of literature to which the findings could be compared and discussed, and therefore, more research is needed to explain this effect.

In Study 6, the findings indicated slightly varied effects of the sterile cockpit rule on performance. The sterile cockpit rule was used to simulate experimental situations without the Background talk factor. Task performance without background talk can be analogous to complying with the sterile cockpit rule and abstaining from pilot–pilot talk. While the effect of background talk was largest in the Call sign recognition task, the differences diminished for Error identification and Prediction tasks. However, it can be argued that the more complicated tasks of comprehension (Error identification task) and prediction (Prediction task) required the participants' full attention, so the performance was not found to be slower or less accurate with simultaneous listening to background talk while performing the tasks, because the background talk was not really attended to. Carretta, Perry and Ree (2009) identified working memory and divided attention to be predictors of SA. This suggests that the need to prioritize between tasks is somewhat natural. This is also consistent with Demany et al. (2010); the more cognitive load the task represents, the less attention is allocated to a simultaneous task. However, there is a lack of available literature to clearly support or refute this interpretation. Moreover, the comparison of studies was not precise and faced several limitations, which were described in the corresponding Discussion sections.

The relevance of the statistically significant or non-significant findings can be discussed according to their potential practical significance. For example, the call sign recognition performance was found to be faster in the pure English condition (0.182 s) when compared to the Mix condition (0.3 s) by 0.118 s. A question is, whether this—relatively small—difference can be practically significant. Only a limited number of radio transmissions in a specified time can be handled in a single radio frequency broadcast (SKYbrary, 2013). The maximum number of transmissions is determined by the length of each transmission and its response (SKYbrary, 2013). With increased number of transmissions, the frequency becomes congested (SKYbrary, 2013). The likelihood of frequency congestion is likely greater in Terminal Control Areas (TMA) with high volume of air traffic. According to Prinzo et al. (2006, p. 9), on average, "one aircraft requested and received air traffic services every 1 min 26 s in the approach sectors and 1 min 6 s in the departure sectors". Moreover, Terminal Control Areas were identified by the participants of Study 1 in relation to adverse effects of language switching on their situation awareness, causing repeated transmissions (see Table 6). In this context, 0.118 s faster responses on transmissions in monolingual as compared to bilingual condition can provide a room for a message that pilots or ATCOs need to call through.

The importance of this finding might also be viewed from another perspective. That is, at the outset of this thesis the answer was not known as to what size effect might be found. For example, had a particularly large effect been detected, then clearly this would have implications. That the effect was small is nevertheless clearly worth discovering.

To sum up, it is a challenge is to put the findings together, given that SA is not a construct divided into three separate levels. Although these findings can provide only limited discussion of the overall SA of bilinguals in monolingual English versus bilingual language environments, it can be assumed that SA will probably not be impaired when air traffic communications are conducted in monolingual English environments. There was some evidence of circumstances in which bilingual pilots might even benefit from monolingual English operations. However, the experimental analyses were limited, as will be discussed in the subsequent section, and as such, no specific or definite assumption can be made.

Instead, it is relevant to consider that the complicacy of the levels of SA increases, but likely not only because they encompass the previous level(s). However, some discrepancies were found in real incident report analysis (e.g., Jones & Endsley, 1996), which were not resolved in this thesis. The overall findings indicate that there is no clear or definitive answer to the central research question.

10.3. Limitations and Future Research

It is important to note the potential limitations in the studies that comprise this thesis. As each study in this thesis outlined specific limitations, only the limitations that apply to the thesis in general are discussed here.

The first limitation relates to the statistical analyses. Non-parametric tests were predominantly used; these tend to be less powerful for detecting an effect than their parametric counterparts. In other words, non-parametric tests are more likely to be susceptible to false negative results (type II error) than their parametric counterparts. It can be assumed that with larger sample sizes, the findings might either reveal an effect that was not detected, or change the observed effect in some way.

The second limitation is related to the measurement of participants' English language proficiency, which was predominantly represented by self-report of the IELTS Listening test score that participants achieved prior to their study. However, while residing in an English-speaking country, their proficiency could have improved. Because the tasks did not put large demands on English language proficiency, more complicated experimental situations involving the use of plain language might reveal some effects. This can be addressed in further research. Additionally, future research can also explore the performance of early bilinguals compared to late bilinguals (Costa & Santesteban, 2004a), as well as those for whom English is not the second language, but rather the third or fourth (e.g., Gabryś-Barker, 2006; Philipp, Gade, & Koch, 2007). An analysis of the effects of second language proficiency and task-induced cognitive workload on participants' speech production and retention of information conducted by Farris et al. (2008) provides more information in the area of English language proficiency in pilot–controller communication.

An issue that has not been addressed in this set of studies is the potentially confounding effect of the use of a computer voice from a text-to-speech programme to record acoustic stimuli. Both experimental stimuli and background talks were recorded using female computerised voices, which could reduce comprehension, especially of Chinese words. Chinese is a tonal language and therefore, the correct intonation is a very important factor. However, the Chinese computerised voice was monotonous. This may have decreased the intelligibility of background talks. More naturalistic pronunciation of Chinese acoustic stimuli should be considered for future research. Another significant limitation of the studies in this thesis related to the stimuli recording was the use of the Audacity programme, which limited the possibility to identify the stimuli decibel levels where no effect of amplify was added. Using the amplify effect allowed setting up the decibel levels by the experimenter. Without using the effect amplify, the SNR was set up by changing the Gain of a signal as compared to noise. However, the Gain feature of Audacity does not show the changes to peak amplitude. The decibel levels of signal and noise could be read out from the amplitude levels, and finally, from the plot spectrum. Both the display of the amplitude levels and the plot spectrum use dB values relative to "full scale" (sometimes written as ‘dbFS’), which is the convention used universally for signal measurement (Steve, 2018). This convention gives negative dB values for all normal signal levels, because all signals recorded in Audacity are with reference to 0dB. Moreover, the decibel levels refer to loudness, which essentially might vary across the participants who were instructed to adjust the volume to the comfortable level in order to prevent any harm to their hearing. This limitation therefore suggests the need to specify the levels of noise and stimuli using more advanced programme for creation of the stimuli, and the need to use a sound level meter to measure the actual loudness of the stimuli during experiment.

It might be argued that all experimental studies (Studies 2–5) used computer-based experiments, which potentially compromises the ecological validity regarding the effects of language switching on SA. This limitation can be subject to empirical confirmation of the language switching tasks being integrated into a more realistic simulation. A flight simulator-based platform would have been preferable, but access to non-native English speaking pilots was almost impossible. Therefore, the findings were obtained from the population of non-aviation students. However, this can also limit generalization of the findings to the population of pilots and ATCOs. Whether similar results would be obtained in actual bilingual air traffic conditions cannot be determined. However, it has been explained that this type of methodology, at the very least, acts as an indicator to real-life behaviour (e.g., Exum, Turner, & Hartman, 2012; Orlady & Orlady, 1999). Therefore, the findings of this thesis should provide some indication of the language switching effect in aviation practice.

The studies in this thesis indicated several areas for further research. Language switching studies, conducted primarily by psycholinguists or neurolinguists, typically identified poorer performance in the language switching condition than in monolingual conditions, with the performance difference termed ‘switch costs’ (e.g., Declerck et al., 2015a, 2012). The language switching paradigm was used in the primary experimental tasks. However, aviation personnel have a multiple-task assignment; the primary task of flying and the secondary task of communicating (Federal Aviation Administration, n.d.). Therefore, there remains a set of unanswered questions about the pilots’ and ATCOs’ language alternation in tasks more related to flying or controlling. This could be further investigated using more realistic experimental settings. Future research can explore language switching as a task secondary to the main task of flying an aircraft.

Bock et al. (2007) conducted a study of the effects of speech production and speech comprehension on driving performance in which simulated driving was the primary task and language-production or language-comprehension was the secondary task. They compared performance under single-task (driving only) and dual-task (driving with a concurrent speech task) conditions. The concurrent speech tasks involved SA, even though the study did not use this terminology. Participants perceived and understood statements about spatial relationships between pairs of buildings on the campus of the University of Illinois, where the simulation experiment took place. Similar studies in aviation could involve analysis of the effect of background talk on flying performance and test the sterile cockpit rule.

Although this research did not precisely follow the language switching paradigm, or typical SA studies (because it was not possible to do so), it can nevertheless be significantly beneficial for future research and practice, given the lack of empirical evidence of effects of language alternation on performance in aviation. The combination of methods is both an advantage and a shortcoming. The primary advantage is that it allowed exploration of the topic from a new perspective, and as such, the potential limitations can facilitate further discussion and investigation. For example, further research can be designed to explore potential intervention steps for improvement of communication in the bilingual air traffic environment. The key aspect to keep in mind when exploring the performance of bilinguals in bilingual versus monolingual air traffic environments is the practical applicability of any of the interventions. Therefore, future research can compare the performance of bilinguals in a bilingual air traffic environment with the performance of monolinguals, who can speak only one of the two languages used in the bilingual airspace and are thus out of the communication loop. By doing so, the potential costs and benefits of suggested interventions could be investigated. The goal is to improve safety, and considerations of the bilingual air traffic environment from various perspectives can only benefit aviation safety.

10.4. Cognitive Implications for Practice

The previous discussion provided a framework for drawing inferences about empirical findings of language switching to aviation practice.

First, the cognitive implications of level 1 SA, recognition of important elements, are reviewed. Besides the implications provided in the Discussion section of Study 2, the findings might generally suggest that people tend to make somewhat hasty decisions based on the minimal information they perceive. Participants' RT increased with increased similarity of two call signs, with responses made even before the entire word was heard. Importantly, this tendency was found in both monolingual English and bilingual condition. The data were interpreted within the RPDM model (Jensen, 2005). In practice, decisions to respond or not respond are made immediately after recognition of a familiar cue. For example, in a flying situation, pilots pick up speech cues (in whichever language) that let them recognise a call sign or a pattern in general. When new information matches the familiar pattern or similar situation stored in memory, the pilots choose an action (to respond or not). This suggests that experience is crucial factor, which is consistent with Carretta, Perry and Ree (2009), who found flying experience to be the best predictor of SA. With respect to communication, a practical implication can lie in a *speed–accuracy trade off* attitude; that is, when there is a tendency to respond immediately, based on the perception of minimal information, aviation personnel may *pause* before deciding, and double-check or listen for more information. Understandably, this attitude can differ depending on the situation; some situations require faster, or even immediate reactions. The speed–accuracy trade off attitude stresses a well-known metaphor ‘measure twice and cut once’, because correcting mistakes might not always be possible.

Next, the finding that performance was more accurate in the L2 condition than in the L1 condition may point to the importance of the frequency with which a second language is used. This finding was attributed to participants residing in an English-speaking country and thus being exposed to English almost all the time. Consequently, increasing exposure to English can improve performance of pilots and ATCOs who do not use English on a radio frequency very often, for example, in countries where international flights are less frequent. Presumably, the more English is used in practice, the more it can facilitate faster and accurate recognisability, because the sensory modality will be adapted to the language. Aviation personnel can benefit from increased use of English to prevent their language skills from deteriorating. The monolingual English air traffic environment can then be proposed not as an obligation, but rather as a benefit for non-native English speakers. Additionally, pilots and ATCOs can listen to radio or television programmes in English during their free time. Importantly, the effect of the sterile cockpit rule was found to be the most significant for level 1 SA, recognition.

The cognitive implications for level 2 SA, based on the Error identification task reported in Study 3, can be reviewed within the practical context of read-backs and automation. The findings of Study 3 suggested that responses to correct information were faster and more accurate than responses to incorrect information. SDT measures further suggested a bias towards *yes* responses; that is, false alarm types of errors. It must be noted that the instruction potentially affected this bias, because participants expected errors to occur. Contrasting this bias with the bias to *no* responses (miss type of errors) observed in the Call sign recognition and Prediction tasks can indicate the potential benefit of expectation. Employing an *attitude of error expectation*, or expecting others to make errors, can potentially mitigate the misdetection of hear-back/read-back errors. With this attitude, pilots and ATCOs can concentrate more attention on what was just transmitted. A similar attitude can have the same effect with automation; that is, expecting that it can fail can lead to checking it more frequently and thoroughly. Of course, such attitudes will not solve all problems and, if not utilized with caution, may bring new problems as well. For example, Dzindolet et al. (2003) found that automation was initially considered trustworthy, but when it sometimes failed participants distrusted even reliable automation. However, distrust and an attitude of error expectation are not necessarily equivalent concepts. Expectation of errors might be proposed as a mindful critical evaluation of a situation with appropriate adjustment of decision criterion, which can be utilised instead of automatic, *a priori* trust/distrust as a reaction based on previous experience but with no consideration for the current circumstances.

It is also important to differentiate between attitude of error expectation and expectation bias in general. Catherwood et al. (2014) explored brain activity during loss of SA associated with the influence of expectation. They suggested that individuals make more errors if situations do not fit expectations. To illustrate this, Catherwood et al. (2014) discussed the crash of a DC10 aircraft operated by Air New Zealand on Mt. Erebus, Antarctica in 1979, where the flight crew's expectation was based on faulty flight path data and might have caused visual information about the location to be overlooked. This example can also serve to support the proposed attitude of error expectation. Speculatively, if the crew had employed an attitude of expecting an error, they might have critically reviewed the flight path data and detected an error. If SA is lost more frequently when a situation does not fit an expectation, then a situation that does not fit an expectation of error should result in safe outcome behaviour.

Importantly, individual differences must be considered prior to adopting any attitude related to the placement of the decision criterion. For example, an attitude to expect an error may not necessarily be helpful for those pilots who tend to apply very liberal criterion; that is, they tend to respond every time there is the slightest indication of an error and have high rates of false alarms in their practice. Moreover, adjustment of the criterion is also largely situationally dependent based on potential consequences; some situations require every indication to be tested because they involve life-or-death issues. The suggestion provided above—adopting an attitude of error expectation—is general and requires individual considerations. In response to this, the importance of Crew Resource Management (CRM) where personal skills are developed and cultivated, is stressed. Individuals can explore their own behavioural biases and adjust the placement of their decision criterion accordingly. Feedback from experienced instructors might be helpful and necessary to guide and facilitate the development of airmanship.

Finally, it is difficult to make any recommendations about level 3 SA, prediction, because it is so hard to measure. The measurement of prediction is relevant to the evaluation of progress, whether in terms of an individual's skills (e.g., CRM training) or in terms of a flight. Further research is necessary to understand and explain prediction in aviation.

In summary, the cognitive implications for practice proposed in this section are intended to raise awareness and influence attitudes rather than provide practical concrete steps of what to do. It was beyond the scope of the current research to propose specific strategies for improving the SA of bilinguals. Increasing awareness was the ultimate goal because the vast variety of possible situations that can be experienced in aviation make it impossible to develop concrete and specific recommendations. Building and applying an awareness-enhancing attitude in individuals can be considered a fruitful and promising start, and can be further expanded by knowledge from future research, CRM training, and practice itself. This leads the discussion to the final chapter, Conclusion.

CONCLUSION

By using a signal detection-based language switching paradigm it was possible to distinguish some effect of different language conditions on cognitive representations of the three levels of situation awareness. Data from language switching studies were taken as evidence that language switching conditions influence the speed and accuracy of performance, by increasing processing time and the number of errors. Even though there was no doubt that language alternation affects performance, it was found in this thesis that the effect differed across the three levels of SA.

Most convincing was the evidence in favour of the monolingual English condition for level 1 SA, recognition. However, with respect to level 2 and 3 SA, the findings showed no evidence of any difference between the monolingual English and bilingual conditions that could be explained within the scope of the current knowledge. Therefore, drawing a conclusion about which language condition facilitates faster and more accurate performance of bilinguals does not appear to be as straightforward as it seemed from the review of aviation accidents involving bilingualism. Although it has not been clearly demonstrated that the actual SA performance changed as a function of language condition, the findings of this thesis can contribute to a theoretical understanding of bilinguals' SA. The three levels of SA were investigated further, with respect to the effect of call sign similarity, error identification and how far ahead it is safe to predict (in terms of performance accuracy).

It was shown that increasing the similarity of two numerical call signs increased both speed and errors of recognition, and the effect appeared to be larger in the Mix condition. Further, detecting an error in the message prolonged the latency in comparison with responding to a correct message. More importantly, the corresponding study revealed an effect of expectation influencing response bias; when errors were expected more false alarm types of errors occurred. In the general discussion, a potential benefit from this effect was proposed. It was also demonstrated that prediction of events or changes in the short term was more accurate than prediction into the more distant future. Findings also supported the importance of the sterile cockpit rule, especially for call sign recognition.

The overall lack of evidence of statistically significant differences between language conditions may indicate that the advantages of bilingualism proposed in previous research are not transferred into improved performance. When considering the out-of-the-loop phenomenon experienced by English-only pilots and ATCOs, which was identified in Study 1, the question arises: Would a universal language for communication on radio frequencies be worth considering, to allow everyone to understand what is said?

A monolingual radiotelephony approach to the use of English that would also include non-commercial aviation may be unachievable (Clark & Tomato, 2017), and therefore, the task remains for future studies to disentangle the effects of language switching on performance of bilingual pilots and ATCOs. Communication errors will probably never be eliminated even if a single language policy were to be implemented. In addition, a one-size policy may not fit all possible conditions or improve safety standards. This leads to the conclusion that focus should be directed toward the individual approach of understanding one's own response biases for the sake of building one's airmanship, given that the safety of the whole system depends on the safe performance of individuals.

REFERENCES

Accident Investigation Board Norway. (2013). Report on supplementary investigation - air accident at Torghatten near Brønnøysund on 6 May 1988 with DHC-7-102, LN-WFN. Lillestrøm: Report SL 2013/29.

Adi-Japha, E., Berberich-Artzi, J., & Libnawi, A. (2010). Cognitive flexibility in drawings of bilingual children. *Child Development, 81*, 1356–1366.

Afshar, P. F., Bahramnezhad, F., Asgari, P., & Shiri, M. (2016). Effect of white noise on sleep in patients admitted to a coronary care. *Journal of Caring Sciences, 5*(2), 103–109.

Agenzia Nazionale per la Sicurezza del Volo. (2004). Final report. Accident involved aircraft Boeing MD-87, registration SE-DMA and Cessna 525-A, registration D-IEVX, Milano, Linate Airport, October 8, 2001. N. A/1/04

Air Force Flight Standards Agency (AFFSA) (1998). Crew resource management (CRM). Basic concepts. *Air Traffic Control Training Series, AT-M-06A*. United States Air Force.

Aircraft Accident Investigation Commission (AAIC). (1976). British Airways Trident G-AWZT, Inex Adria DC9 YU-AJR: Report on the collision in the Zagreb area, Yugoslavia, on 10 September 1976. *Aircraft Accident Report 5/77*. London: Her Majesty's Stationery Office.

Aitchison, J. (1994). *Words in the mind: An introduction to the mental lexicon* (2nd ed.). Oxford: Blackwell.

Alcock, C. (2007). Language confusion led to fatal incursion. *Aviation International News Online*. Retrieved from https://www.ainonline.com/aviation-news/2007-11-27/language-confusion-led-fatal-incursion.

Allendoerfer, K. R., Pai, S., & Friedman-Berg, F. J. (2008). The complexity of signal detection in air traffic control alert situations. *Proceedings of the Human Factors and Ergonomics Society, 52nd Annual Meeting, 52(1), 54–58.*

Anthony, K., Wiencek, C., Bauer, C., Daly, B., & Anthony, M. K. (2010). No interruptions please: Impact of a no interruption zone on medication safety in intensive care units. *Critical Care Nurse, 30*(3), 21–29.

Arrabito, G. R. (2009). Effects of talker sex and voice style of verbal cockpit warnings on performance. *Human Factors, 51*(1), 3–20.

Aryadoust, V. (2013). *Building a validity argument for a listening test of academic proficiency*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Atkinson, R. C. & Shiffrin, R. M. (1971). *The control processes of short-term memory*. Technical Report: Institute for Mathematical Studies in the Social Sciences, Stanford University.

Auton, J. C., Wiggins, M. W., Searle, B. J., & Rattanasone, N. X. (2016). Utilization of prosodic and linguistic cues during perceptions of nonunderstandings in radio communication. *Applied Psycholinguistics, 38*(3), 509–539.

Aviation Safety Reporting System [ASRS] Report No. 995712 (2012). Retrieved from https://akama.arc.nasa.gov/ASRSDBOnline/QueryWizard_Display.aspx?exportToWord=Y&server=ASRSO.

Avery, P. & Ehrlich, S. L. (1992). *Teaching American English pronunciation.* New York: Oxford University Press.

Aviation Safety Network (n.d.). *Aircraft Ilyushin 76TD UN-76435 Charkhi Dadri*. Accident report. Retrieved from https://aviation-safety.net/database/record.php?id=19961112-1.

Baddeley, A. (1997). *Human memory: Theory and practice.* Erlbaum: Psychology Press.

Badger, R. & Yan, X. (2009). The use of tactics and strategies by Chinese students in the Listening component of IELTS. *IELTS Research Reports, 9.*

Badger, R. & Yan, X. (2012). To what extent is communicative language teaching a feature of IELTS classes in China? *IELTS Research Reports*, *13*.

Banbury, S. P., Croft, D. G., Macken, W. J., & Jones, D. M. (2004). A cognitive streaming account of situation awareness. In S. Banbury & S. Tremblay (Eds.), *A cognitive approach to situation awareness: Theory and application* (pp. 117–134). Burlington, VT: Ashgate.

Baron, R. A. (1995). *The cockpit, the cabin, and social psychology*. Retrieved from https://web.archive.org/web/20131204222509/https://airlinesafety.com/editorials/CockpitCabinPsychology.htm.

Barshi, I. & Farris, C. (2013). *Misunderstandings in ATC communication: Language, cognition, and experimental methodology*. Burlington, VT: Ashgate.

Bhatia, T. K. (2017). *Bilingualism and multilingualism*. Retrieved from http://www.oxfordbibliographies.com/view/document/obo-9780199772810/obo-9780199772810-0056.xml.

Bhatia, T. K. & Ritchie, W. C. (2013). *The handbook of bilingualism and multilingualism* (2nd ed.). Chichester, United Kingdom: Blackwell Publishing, Ltd.

Bialystok, E. (2009). Bilingualism: The good, the bad and the indifferent. *Bilingualism: Language and Cognition, 12*(3), 11.

Bialystok, E. (2010). Global–local and trail-making tasks by monolingual and bilingual children. *Developmental Psychology, 46*, 93–105.

Bialystok, E., Craik, F. I. M., & Luk, G. (2012). Bilingualism: Consequences for mind and brain. *Trends in Cognitive Sciences, 16*, 240–250.

Billings, C. E. (1995). Situation awareness measurement and analysis: A commentary. *Proceedings of the International Conference on Experimental Analysis and Measurement of Situation Awareness*. Embry-Riddle Aeronautical University Press, FL.

Billings, C. E. & Cheaney, E. S. (1981). *Information transfer problems in the aviation system*. Technical Paper 1875. NASA Ames Research Center: Washington, D. C.

Blanchet, D. (2017). Sterile cockpit, sterile crew. *EMS World*. Retrieved from https://www.emsworld.com/article/219018/sterile-cockpit-sterile-crew.

Bobb, S. C. & Wodniecka, Z. (2013). Language switching in picture naming: What asymmetric switch costs (do not) tell us about inhibition in bilingual speech planning. *Journal of Cognitive Psychology, 25*(5), 568–585.

Bock, K., Dell, G. S., Garnsey, S. M., Kramer, A. F., & Kubose, T. T. (2007). Car talk, car listen. In A. S. Meyer, L. R. Wheeldon, & A. Krott (Eds.), *Automaticity and control in language processing* (pp. 21–42). New York: Psychology Press Taylor & Francis Group.

Borins, S. F. (1983). *The language of the skies: The bilingual air traffic control conflict in Canada*. Kingston: Institute of Public Administration of Canada.

Boudes, N. & Cellier, J.-M. (2000). Accuracy of estimations made by air traffic controllers. *International Journal of Aviation Psychology*, *10*(2), 207–225.

Branzi, F. M., Calabria, M., Boscarin, M. L., & Costa, A. (2016). On the overlap between bilingual language control and domain-general executive control. *Acta Psychologica*, *166*, 21–30.

Breeze, R. & Miller, P. (2011). Predictive validity of the IELTS Listening Test as an indicator of student coping ability in Spain. *IELTS Research Reports*, *12*.

British Council (n.d.a). *Understand the IELTS test format*. Retrieved from http://takeielts.britishcouncil.org/prepare-test/understand-test-format.

British Council (n.d.b). *Understand how to calculate your IELTS scores*. Retrieved from http://takeielts.britishcouncil.org/find-out-about-results/understand-your-ielts-scores.

British Council (n.d.c). *Understand the Listening Test.* Retrieved from https://takeielts.britishcouncil.org/prepare-test/understand-test-format/listening-test.

British Council (n.d.d). *Listening practice test 1*. Retrieved from https://takeielts.britishcouncil.org/prepare-test/free-ielts-practice-tests/listening-practice-test-1.

British Council (n.d.e). *IELTS Test Report Form*. Retrieved from https://takeielts.britishcouncil.org/find-out-about-results/results-process/test-report-form.

British Council (2018). *Why choose IELTS?* Retrieved from http://www.britishcouncil.org.ua/en/exam/ielts/why-choose.

Bryant, D. J., Lichacz, F. M. J., Hollands, J. G., & Baranski, J. V. (2004). Modelling situation awareness in an organizational context: Military command and control. In S.

Banbury, & S. Tremblay (Eds.), *A cognitive approach to situation awareness: theory and application* (pp. 104–116). Burlington, VT: Ashgate.

Bubic, A., von Cramon, D. Y., & Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in Human Neuroscience*, *4*(25), 1–15.

Buchanan, T. W., Laures-Gore, J. S., & Duff, M. C. (2014). Acute stress reduces speech fluency. *Biological Psychology, 97*, 60–66.

Bultena, S., Dijkstra, T., & van Hell, J. G. (2015). Switch cost modulations in bilingual sentence processing: Evidence from shadowing. *Language, Cognition and Neuroscience*, *30*(5), 586–605.

Bureau d'Enquêtes et d'Analyses (2000). *Accident on 25 May 2000 at Paris Charles de Gaulle (95) to aircraft F-GHED operated by Air Liberté and G-SSWN operated by Streamline Aviation*. Report BEA-F-ED000525. Retrieved from https://reports.aviation-safety.net/2000/20000525-0_SH33_G-SSWN.pdf.

Bureau d'Enquêtes et d'Analyses (2016). *Final Investigation Report: Accident to the Airbus A320-211, registered D-AIPX and operated by Germanwings, flight GWI18G, on 03/24/15 at Prads-Haute-Bléone*. Report BEA2015-0125. Retrieved from https://www.bea.aero/uploads/tx_elydbrapports/BEA2015-0125.en-LR.pdf.

Burns, J. F. (1996, November 13). Two airliners collide in midair, killing all 351 aboard in India. *The New York Times*. Retrieved from https://www.nytimes.com/1996/11/13/world/two-airliners-collide-in-midair-killing-all-351-aboard-in-india.html?pagewanted=print&src=pm.

Cambridge English (2016). *Comparing scores to IELTS: Cambridge English: Advanced (CAE) and Cambridge English: First (FCE)*. University of Cambridge: Cambridge English Language Assessment.

Campbell, J. I. D. (2005). Asymmetrical language switching costs in Chinese–English bilinguals' number naming and simple arithmetic. *Bilingualism: Language & Cognition*, *8*(1), 85–91.

Campbell, R. D. & Bagshaw, M. (2002). *Human performance and limitations in aviation* (3rd ed.). London: Blackwell Science Ltd.

Canadian Encyclopedia (n.d.). *Official Languages Act (1969)*. Retrieved from http://www.thecanadianencyclopedia.ca/en/article/official-languages-act-1969/.

Cardosi, K. (1993). *An analysis of en route controller–pilot voice communications*. No. DOT/FAA/RD-93/11. Washington, DC: Federal Aviation Administration.

Carretta, T. S., Perry, D. C., & Ree, M. J. (2009). Prediction of situational awareness in F-15 pilots. *International Journal of Aviation Psychology*, *6*(1), 21–41.

Catherwood, D., Edgar, G. K., Nikolla, D., Alford, C. H., Brookes, D., Baker, S., & White, S. (2014). Mapping brain activity during loss of situation awareness: An EEG investigation of a basis for top-down influence on perception. *Human Factors*, *56*(8), 1428–1452.

Chan, J. W. & Simpson, C. A. (1990). *Comparison of speech intelligibility in cockpit noise using SPH-4 helmet with and without active noise reduction*. National Aeronautics and Space Administration [NASA] Contractor Report 177564. California: Ames Research Center.

Chaplin, J. P. (1985). *Dictionary of psychology*. New York, NY: Laurel.

Chen, C. H. (2014). A contrastive study of time as space metaphor in English and Chinese. *Theory and Practice in Language Studies*, *4*(1), 129–136.

Chen, S., Kopald, H., Chong, R. S., Wei, Y-J., & Levonian, Z. (2017). Readback error detection using automatic speech recognition. *Twelfth USA/Europe Air Traffic Management Research and Development Seminar*. Retrieved from file:///E:/desktop_new_lit_raw%20material/12th_ATM_RD_Seminar_paper_20.pdf.

Cheng, Y.L. & Howard, D. (2008). The time cost of mixed-language processing: an investigation. *International Journal of Bilingualism*, *12*(3), 209–222.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, *25*, 975–979.

Chincotta, D. & Underwood, G. (1997). Digit span and articulatory suppression: A cross-linguistic comparison. *European Journal of Cognitive Psychology*, *9*(1), 89–96.

Christoffels, I. K., Firk, C., & Schiller, N. O. (2007). Bilingual language control: An event-related brain potential study. *Brain Research*, *1147*, 192–208.

Christoffels, I. K., Kroll, J. F., & Bajo, M. T. (2013). Introduction to bilingualism and cognitive control. *Frontiers in Psychology*, *4*, 199.

Chung-Fat-Yim, A., Sorge, G. B. & Bialystok, E. (2017). The relationship between bilingualism and selective attention in young adults: Evidence from an ambiguous figures task. *Quarterly Journal of Experimental Psychology*, *70*(3), 366–372.

Chute, R. D. & Wiener, E. L. (1996). Cockpit–cabin communication: II. Shall we tell the pilots? *International Journal of Aviation Psychology, 6*(3), 211–231.

Civil Aviation Authority (2000). *Aircraft Call Sign Confusion Evaluation Safety Study [ACCESS], CAP 704.* West Sussex: UK Civil Aviation Authority.

Civil Aviation Authority (2014). *Flight-crew human factors handbook, CAP 737*. West Sussex: UK Civil Aviation Authority.

Clark, B. & Tomato, Y. S. (2017). *Aviation English research project: Data analysis findings and best practice recommendations*. West Sussex: UK Civil Aviation Authority.

Coleman, G. & Heap, S. (1998). The misinterpretation of directions for the questions in the academic reading and listening sub-tests of the IELTS test. *IELTS Research Reports, 1.*

Comisión de Investigación de Accidentes e Incidentes de Aviación Civil. (2013). Report IN-007/2012. *Addenda Bulletin 1/2013*. Retrieved from https://www.skybrary.aero/bookshelf/books/2290.pdf.

Conrad, R. (1964). Acoustic confusions in immediate memory. *British Journal of Psychology*, *55*(1), 75–84.

Cookson, S. (2009). Zagreb and Tenerife. Airline accidents involving linguistic factors. *Australian Review of Applied Linguistics*, *32*(3), 22.1–22.14.

Corradini, P. & Cacciari, C. (2002). The effect of workload and workshift on air traffic control: A taxonomy of communicative problems. *Cognition, Technology & Work, 4*, 229–239.

Costa, A., Hernández, M., Costa-Faidella, J., & Sebastián-Gallés, N. (2009). On the bilingual advantage in conflict processing: Now you see it, now you don’t. *Cognition*, *113*(2), 135–149.

Costa, A., Hernández, M., & Sebastián-Gallés, N. (2006). Bilingualism aids conflict resolution: Evidence from the ANT task. *Cognition*, *106*, 59–86.

Costa, A. & Miozzo, M., & Caramazza, A. (1999). Lexical selection in bilinguals: Do words in the bilingual's two lexicons compete for selection? *Journal of Memory & Language*, *41*, 365–397.

Costa, A. & Santesteban, M. (2004a). Lexical access in bilingual speech production: Evidence from language switching in highly proficient bilinguals and L2 learners. *Journal of Memory & Language*, *50*, 491–511.

Costa, A. & Santesteban, M. (2004b). Bilingual word perception and production: Two sides of the same coin? *Trends in Cognitive Sciences*, *8*(6), 253.

Costa, A., Santesteban, M., & Ivanova, I. (2006). How do highly proficient bilinguals control their lexicalization process? Inhibitory and language-specific selection mechanisms are both functional. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *32*, 1057–1074.

Cox, S. & Vinagre, L. (2004). Modelling of confusions in aircraft call-sings. *Speech Communication*, *42*, 289–312.

Crew Resource Management. (2018). *Attention and perception*. Retrieved from http://www.crewresourcemanagement.net/information-processing/attention-and-perception.

Cramer, D. (1998). *Fundamental statistics for social research: Step-by-step calculations and computer techniques using SPSS for Windows*. New York: Routledge.

Croft, D. G., Banbury, S. P., Butler, L. T., & Berry, D. C. (2004). The role of awareness in situation awareness. In S. Banbury & S. Tremblay (Eds.), *A cognitive approach to situation awareness: theory and application* (pp. 82–103). Burlington, VT: Ashgate.

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

Cushing, S. (1995). Pilot–air traffic control communications: It's not (only) what you say, it's how you say it. *Flight Safety Digest*, *14*(7), 1–10.

Daniel, J. (1984). *Psychická záťaž v laboratórnych a terénnych podmienkach. [Psychological load in laboratory and terrain conditions]*. Bratislava: Veda.

Daniel, J. & Pikala, I. (1976). *Psychológia práce. [Psychology of work]*. Bratislava: Práca.

David, O. A., Matu, S. A., Pintea, S., Cotet, C. D., & Nagy, D. (2014). Cognitive-behavioral processes based on using the ABC analysis by trainees for their personal development. *Journal of Rational-Emotive & Cognitive-Behavior Therapy, 32*, 198–215.

Declerck, M., Grainger, J., Koch, I., & Philipp, A. M. (2017). Is language control just a form of executive control? Evidence for overlapping processes in language switching and task switching. *Journal of Memory & Language, 95*, 138–145.

Declerck, M., Koch I., & Philipp, A.M. (2012). Digits vs. pictures: The influence of stimulus type on language switching. *Bilingualism: Language & Cognition*, *15*(4), 896–904.

Declerck, M., Koch, I., & Philipp, A. M. (2015). The minimum requirements of language control: Evidence from sequential predictability effects in language switching. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *41*(2), 377–394.

Declerck, M. & Philipp, A. M. (2015a). A sentence to remember: Instructed language switching in sentence production. *Cognition*, *137*, 166–173.

Declerck, M. & Philipp, A. M. (2015b). A review of control processes and their locus in language switching. *Psychonomic Bulletin & Review*, *22*(6), 1630–1645.

Declerck, M., Philipp, A. M., & Koch, I. (2013). Bilingual control: Sequential memory in language switching. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *39*(6), 1793–1806.

Declerck, M., Stephan, D. N., Koch, I., & Philipp, A. M. (2015). The other modality: Auditory stimuli in language switching. *Journal of Cognitive Psychology*, *27*, 685–691.

Defense Language Institute (1974). *A contrastive study of English and Mandarin Chinese*. Monterey, CA: Defense Language Institute

Dehaene, S. & Cohen, L. (1997). Cerebral pathways for calculation: Double dissociation between rote verbal and quantitative knowledge of arithmetic. *Cortex*, *33*, 219–250.

Dekker, S. W. A. (2003). Illusions of explanation: A critical essay on error classification. *International Journal of Aviation Psychology*, *13*(2), 95–106.

Demany, L., Semal, C., Cazalets, J-R, & Pressnitzer, D. (2010). Fundamental differences in change detection between vision and audition. *Experimental Brain Research*, *203*(2), 261–270.

Dennis, W. (2015a, March 2). Airlines complain about ATC's use of Mandarin at Beijing. *Aviation International News Online.* Retrieved from http://www.ainonline.com/aviation-news/air-transport/2015-03-02/airlines-complain-about-atcs-use-mandarin-beijing.

Dennis, W. (2015b, November 26). Chinese ATC to adopt English-only policy. *Aviation International News Online.* Retrieved from http://www.ainonline.com/aviation-news/air-transport/2015-11-26/chinese-atc-adopt-english-only-policy.

Denti (2011, July 11). Re: European call signs [Online forum comment]. Retrieved from http://www.pprune.org/tech-log/455205-european-call-signs.html.

Drullman, R. & Bronkhorst, A. (2000). Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation. *Journal of the Acoustical Society of America*, *107*, 2224–2235.

Durso, F. & Gronlund, S. D. (1999). Situation awareness. In F. T. Durso, R. Nickerson, R. Schvaneveldt, S. Dumais, M. Chi, S. & Lindsay (Eds.), *Handbook of applied cognition* (pp. 283–314). New York: Wiley.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718.

Edgar, G. K., Catherwood, D., Baker, S., Sallis, G., Bertels, M., Edgar, H. E., & Whelan, A. (2018). Quantitative Analysis of Situation Awareness (QASA): modelling and measuring situation awareness using signal detection theory. *Ergonomics*, *61*(6), 762–777.

Edworthy J, Hellier E, & Rivers J. (2003). The use of male or female voices in warnings systems: A question of acoustics. *Noise Health, 6*, 39-50.

Egan, J. P. (1975). *Signal detection theory and ROC-analysis*. New York: Academic Press.

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. *Proceedings of the Human Factors Society 32nd Annual Meeting* (pp. 97–101). Santa Monica, CA: Human Factors and Ergonomics Society.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*(1), 32–64.

Endsley, M. R. (2000). Theoretical underpinnings of situation awareness: A critical review, In M. R. Endsley & D. J. Garland (Eds.), *Situation awareness analysis and measurement* (pp. 3–32). Mahwah, NJ: L. Erlbaum Associates.

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behaviour Research Methods, Instruments, & Computers*, *28*(1), 1–11.

Ericson, M. A. & McKinley, R. L. (2001). *The intelligibility of multiple talkers separated spatially in noise*. *Final report for the period December 1987 to September 1993*. United States Air Force Research Laboratory.

Estival, D., Farris, C., & Molesworth, B. (2016). *Aviation English: A lingua franca for pilots and air traffic controllers*. Milton Park: Routledge.

Estival, D. & Molesworth, B. (2009). A study of EL2 pilots radio communication in the general aviation environment. *Australian Review of Applied Linguistics, 32*, 24.1–24.16.

Estival, D. & Molesworth, B. (2012). Radio miscommunication: EL2 pilots in the Australian general aviation environment. *Linguistics & the Human Sciences, 5*(3), 351–378.

Etem, K. & Patten, M. (1998). Communications-related incidents in general aviation dual flight training. *ASRS Directline*, 10. Retrieved from https://asrs.arc.nasa.gov/publications/directline/dl10_gacom.htm.

EUROCONTROL. (2006). *Air-Ground Communications Safety Study: Causes and Recommendations*. Retrieved from https://www.skybrary.aero/bookshelf/books/162.pdf.

EUROCONTROL. (2012). *Call sign confusion—Eurocontrol's call sign similarity tool helps improve flight safety.* Retrieved from http://www.eurocontrol.int/news/call-sign-confusion-eurocontrols-call-sign-similarity-tool-helps-improve-flight-safety.

European Aviation Safety Agency. (2012). *Regulation (EU) No 965/2012 on Air Operations, Annex I (Definitions) and Annex III (Part ORO)*. Retrieved from https://www.easa.europa.eu/faq/19134.

Exam English Ltd. (n.d.). *IELTS*. Retrieved from https://www.examenglish.com /IELTS/.

Exum, M. L., Turner, M. G., & Hartman, J. l. (2012). Self-reported intentions to offend: All talk and no action? *American Journal of Criminal Justice*, *37*(4), 523–543.

Fabbro, F. (1999). *The neurolinguistics of bilingualism: An introduction*. Hove: Psychology Press.

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

Federal Aviation Administration. (2008). *Aeronautical decision making*. *P-8740-69.* Retrieved from https://www.faasafety.gov/files/gslac/library/documents/2011/Aug/56413/FAA%20P-8740-69%20Aeronautical%20Decision%20Making%20[hi-res]%20branded.pdf.

Federal Aviation Administration. (2012). *Instrument flying handbook*. *FAA-H-8083-15B.* U. S. Department of Transportation.

Federal Aviation Administration. (2015). *Order JO 7110.65W. Subject: Air Traffic Control*. Retrieved from https://www.faa.gov/documentLibrary/media/Order/ATC.pdf.

Federal Aviation Administration. (n.d.). Fly the aircraft first. *FAA Aviation Safety Briefing*. Retrieved from https://www.faa.gov/news/safety_briefing/2015/media/SE_Topic_15_01.pdf.

Federal Aviation Regulations (1981). *14 CFR 121.542 – Flight Crew Member Duties*. Doc. No. 20661, 46 FR 5502. Retrieved from https://www.gpo.gov/fdsys/granule/CFR-2011-title14-vol3/CFR-2011-title14-vol3-sec121-542/content-detail.html.

Federwisch, M., Ramos, H., & Adams, S. C. (2014). The sterile cockpit: An effective approach to reducing medication errors? How one nursing unit tried to limit interruptions during medication administration by adapting the aviation industry rule. *American Journal of Nursing, 114*(2), 47–55.

Festman, J., Rodriguez-Fornells, A., & Münte, T. F. (2010). Individual differences in control of language interference in late bilinguals are mainly related to general executive abilities. *Behavioral & Brain Functions, 6*, article 5.

Field, A. (2012). *Repeated measures ANOVA. Discovering statistics*. Retrieved from http://www.discoveringstatistics.com/docs/repeatedmeasures.pdf.

Field, A. (2017). *Misconception Mutt extract from chapter 15. Discovering statistics using IBM SPSS statistics*. Retrieved from https://edge.sagepub.com/field5e/chapter-specific-resources/15-glm-4-repeated-measures-designs/misconception-mutt-extract.

FitzPatrick, I. (2011). *Lexical interactions in non-native speech comprehension: Evidence from electro-encephalography, eye-tracking, and functional magnetic resonance imaging* (PhD thesis). Nijmegen, Radboud University Nijmegen.

Flight Safety Digest (1992, August). Cockpit chatter leads to crash. *Flight Safety Foundation* (p. 18). Retrieved from https://flightsafety.org/fsd/fsd_aug92.pdf.

Flin, R., O'Connor, P., & Crichton, M. (2008). *Safety at the sharp end: A guide to non-technical skills*. Aldershot, UK: Ashgate.

Ford, C. (2015, August 26). *Understanding Q-Q Plots*. Retrieved from https://data.library.virginia.edu/understanding-q-q-plots/.

Ford, J., Henderson, R., & O'Hare, D. (2013). Barriers to intra-aircraft communication and safety: The perspective of the flight attendants. *International Journal of Aviation Psychology, 23*(4), 368–387.

Foucart, A., Martin, C. D., Moreno, E. M., & Costa, A. (2014). Can bilinguals see it coming? Word anticipation in L2 sentence reading. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 40*(5), 1–9.

Frankfurt International School (n.d.). *Introduction to language differences*. Retrieved from http://esl.fis.edu/grammar/langdiff/intro.htm#info.

Friedman, E. M. (n.d.). *A commentary to outliers: To drop or not to drop by Grace-Martin, K.* Retrieved from https://www.theanalysisfactor.com/outliers-to-drop-or-not-to-drop/.

Friesen, D. C., Latman, V., Calvo, A., & Bialystok, E. (2015). Attention during visual search: The benefit of bilingualism. *International Journal of Bilingualism*, *19*, 693–702.

Gabryś-Barker, D. (2006). Language activation in the thinking processes of a multilingual language user. *International Journal of Multilingualism*, *3*(2), 105–124.

Gaëtan, T. (2016, July 31). Live ATC: Paris / Roissy CDG Approach / LFPG [Audio file]. Retrieved from https://www.youtube.com/watch?v=3M7dc-tQE1s&t=2s.

Gat, I. B. & Keith, R. W. (1978). An effect of linguistic experience. Auditory word discrimination by native and non-native speakers of English. *Audiology, 17*, 339–345.

Ge, J., Peng, G., Lyua, B., Wanga, Y., Zhuoe, Y., Niuf, Z., & Gaoa, J.-H. (2015). Cross-language differences in the brain network subserving intelligible speech. *Proceedings of the National Academy of Sciences of the United States of America (PNAS), 112*(10), 2972–2977.

Geary, D. C., Cormier, P., Goggin, J. P., Estrada, P., & Lunn, M. C. E. (1993). Mental arithmetic: A componential analysis of speed-of-processing across monolingual, weak bilingual, and strong bilingual adults. *International Journal of Psychology*, *28*(2), 185–201.

Gelfand, S. A. (1998). *Hearing: An introduction to psychological and physiological acoustics*. New York: Marcel Dekker.

George, D. & Mallery, M. (2010). *SPSS for Windows step by step: A simple guide and reference (17.0 update)*. Boston: Pearson.

Girden, E. (1992). *ANOVA: Repeated measures*. Newbury Park, CA: Sage.

Gladwell, M. (2008). *Outliers: The story of success*. New York: Little, Brown & Company.

Gollan, T. H. & Goldrick, M. (2016). Grammatical constraints on language switching: Language control is not just executive control. *Journal of Memory and Language*, *90*, 177–199.

Gollan, T. H., Sandoval, T., & Salmon, D. P. (2011). Cross-language intrusion errors in aging bilinguals reveal the link between executive control and language selection. *Psychological Science*, *22*(9), 1155–1164.

Gollan, T. H., Schotter, E. R., Gomez, J., Murillo, M., & Rayner, K. (2014). Multiple levels of bilingual language control: Evidence from language intrusions in reading aloud. *Psychological Science*, *25*(2). 585–595.

Gordon, P. (2004). Numerical cognition without words: Evidence from Amazonia. *Science, 306,* 496–499.

Grabner, R. H., Saalbach, H., & Eckstein, D. (2012). Language-switching costs in bilingual mathematics learning. *Mind, Brain, and Education*, *6*(3), 147–155.

Grady, M. (2017, October 19). *ERAU investigates language as safety issue*. Retrieved from https://www.avweb.com/avwebflash/news/ERAU-Investigates-Language-As-Safety-Issue-229799-1.html.

Grasu, D. (2015). *Tonal vs. non-tonal languages: Chinese vs. English*. Retrieved from http://www.lexington.ro/en/blog/item/29-tonal-vs-non-tonal-languages-chinese-vs-english.html.

Grayson, R. L. & Billings, Ch. E. (1981). Information transfer between air traffic control and aircraft: Communication problems in flight operations. In C. E. Billings, & E. S. Cheaney (Eds.), *Information transfer problems in the aviation system* (pp. 47–61). Washington, D. C. Ames Research Center, NASA TP-1875.

Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language & Cognition*, *1*, 67–81.

Green, D. M. & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: John Wiley & Sons, Inc.

Green, D. M. & Swets, J. A. (1988). *Signal detection theory and psychophysics*. Revised edition. Los Altos, California: Peninsula.

Grosjean, F. (1982). *Life with two languages: An introduction into bilingualism*. Cambridge, MA: Harvard University Press.

GRSites (n.d.). Free Sound Effects FX Library. aircraft057.wav - Propeller plane - interior sound in-flight [Audio file]. Retrieved from http://www.grsites.com/archive/sounds/ category/21/?offset=48.

Guest, G., MacQueen, K. M., & Namey, E. E. (2012). *Applied thematic analysis.* Thousand Oaks, CA: Sage.

Habrat, A. (2013). The effect of affect on learning: self-esteem and self-concept. In E. Piechurska-Kuciel, & E. Szymanska-Czaplak (Eds.), *Language in cognition and affect* (pp. 239–253). Heidelberg: Springer-Verlag.

Hall, R. (2017, May 30). *Can Chinese be airline pilots?* Retrieved from https://www.quora.com/Can-Chinese-be-airline-pilots.

Hanulova, J., Davidson, D. J., & Indefrey, P. (2011). Where does the delay in L2 picture naming come from? Psycholinguistic and neurocognitive evidence on second language word production. *Language and Cognitive Processes, 26*(7), 902–934.

Harcourt Assessment Inc. (2006). *Predicting ICAO Levels from Versant for English*. Retrieved from https://www.pearsonassessments.com/hai/images/dotcom/vaet/ICAO PredictionFromVersant.pdf.

Heeger, D. (2003). *Signal detection theory*. Department of Psychology, New York University. Retrieved from http://www.cns.nyu.edu/~david/handouts/sdt/sdt.html.

Henneberger, M. (2001). October 7–13; Planes Collide, Killing 118. *The New York Times*. Retrieved from https://www.nytimes.com/2001/10/14/weekinreview/october-7-13-planes-collide-killing-118.html.

Heredia, R. R. & Altarriba, J. (2001). Bilingual language mixing: Why do bilinguals code-switch? *Current Directions in Psychological Science*, *10,* 164–168.

Hermans, D., Bongaerts, T., De Bot, K., & Schreuder, R. (1998). Producing words in a foreign language: Can speakers prevent interference from their first language? *Bilingualism: Language & Cognition*, *1*(3), 213–229.

Hermans, D., Ormel, E., van Besselaar, R., & van Hell, J. (2011). Lexical activation in bilinguals' speech production is dynamic: How language ambiguous words can affect cross-language activation. *Language & Cognitive Processes*, *26*(10), 1687–1709.

Hernandez, M., Costa, A., Fuentes, L. J., Vivas, A. B., & Sebastián-Gallés, N. (2010). The impact of bilingualism on the executive control and orienting networks of attention. *Bilingualism: Language & Cognition*, *13*(3), 315–325.

Hiscock, M., Inch, R., & Kinsbourne, M. (1999). Allocation of attention in dichotic listening: Differential effects on the detection and localization of signals. *Neuropsychology*, *13*(3), 404–414.

Hoaglin, D. C. & Iglewitcz, B. (1987). Fine tuning some resistant rules for outliers labelling. *Journal of American Statistical Association*, *82,* 1147–1149.

Hodgetts, H., Farmer, E., Joose, M., Parmentier, F., Schaefer, D., Hoogeboom, P., & Jones, D. (2005). The effects of party line communication on flight task performance, In D. de Waard, K. A. Brookhuis, R. van Egmond, & T. Boersema (Eds.), *Human factors in design, safety, and management* (pp. 1–12). Maastricht: Shaker Publishing.

Hoekstra, R., Kiers, H. A. L. & Johnson, A. (2012). Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in Psychology, 3*(137). 1–9.

Hoffman, R. (2015). Origins of situation awareness: Cautionary tales from the history of concepts of attention. *Journal of Cognitive Engineering & Decision Making, 9*(1), 73–83.

Hohenhaus, S. M. & Powell, S. M. (2008). Distractions and interruptions: Development of a healthcare sterile cockpit. *Newborn & Infant Nursing Review, 8*(2), 108–110.

Holcomb, P. J. & Neville, H. J. (1991). The electrophysiology of spoken sentence processing. *Psychobiology*, *19,* 286–300.

Hradecky, S. (2013). Incident: Brussels A319 and Iberia A320 at Barcelona on Feb 8th 2012, loss of separation on final approach. *The Aviation Herald.* Retrieved from http://avherald.com/h?article=44c5f743/0000&opt=0.

IBM Knowledge Centre (n.d.). *PLOT Subcommand (EXAMINE command).* Retrieved from https://www.ibm.com/support/knowledgecenter/en/SSLVMB_25.0.0/statistics_reference_project_ddita/spss/base/syn_examine_plot.html.

Ibrahim, R., Shoshani, R., Prior, A., & Share, D. (2013). Bilingualism and measures of spontaneous and reactive cognitive flexibility. *Psychology*, *4*(7A), 1–10.

Ifrah, G. (2000). *The universal history of numbers: From prehistory to the invention of the computer*. New York: John Wiley & Sons.

International Air Transport Association. (2014). *Aviation benefits beyond borders*, 2–8. Retrieved from www.aviationbenefitsbeyondborders.org.

International Civil Aviation Organization. (2001a). *Annex 10 to the Convention on International Civil Aviation. Aeronautical Telecommunications, Volume II Communication Procedures including those with PANS status* (6th ed.). Montreal: ICAO.

International Civil Aviation Organization. (2001b). *Annex 11 to the Convention on International Civil Aviation. Air Traffic Services. Air Traffic Control Service, Flight Information Service, Alerting Service* (13th ed.). Montreal: ICAO.

International Civil Aviation Organization. (2010). *Manual on the implementation of ICAO language proficiency requirements*. *Doc 9835/AN453* (2nd ed.). Montreal: ICAO.

International Civil Aviation Organization. (2011). *Annex 1 to the Convention on International Civil Aviation. Personnel Licensing* (11th ed.). Montreal: ICAO.

International Civil Aviation Organization. (2013). *ICAO announces revamped aviation English language test service site*. Retrieved from https://www.icao.int/Newsroom/Pages/ICAO-announces-revamped-aviation-english-language-test-service-site.aspx.

International Civil Aviation Organization. (2015). *Guidance material related to call sign similarity*. *RASG-MID Safety Advisory – 04 (RSA-04)*. Ref. No. RASG-MID/CSC/01.

International Civil Aviation Organization. (2016). *Designators for Aircraft Operating Agencies, Aeronautical Authorities and Services*. *Doc 8585/176*. Montreal: ICAO.

International Civil Aviation Organization. (n.d.a). *History*. Retrieved from https://www.icao.int/secretariat/TechnicalCooperation/Pages/history.aspx.

International Civil Aviation Organization. (n.d.b). *ICAO Recognized Tests*. Retrieved from https://www4.icao.int/aelts/Home/RecognizedTests.

IELTS (n.d.a). *Research reports*. Retrieved from https://www.ielts.org/teaching-and-research/research-reports.

IELTS (n.d.b). *Test statistics*. Retrieved from https://www.ielts.org/teaching-and-research/test-statistics.

IELTS-up (n.d.). *IELTS Listening Practice Tests*. Retrieved from http://ielts-up.com/listening/ielts-listening-practice.html.

Ison, D. C. (2011). An analysis of statistical power in aviation research. *International Journal of Applied Aviation Studies*, *11*(1), 67–84.

Jackson, L., Chapman, P., & Crundall, D. (2009). What happens next? Predicting other road users' behaviour as a function of driving experience and processing time. *Ergonomics*, *52*(2), 154–164.

Jensen, R.S., Guilke, J., & Tigner, R. (2005). Understanding expert aviator judgement. In R. Flin, E. Salas, M. Strub, & L. Martin (Eds.), *Decision making under stress. Emerging themes and applications* (pp. 233–242). Aldershot, Ashgate Publishing Limited.

Jones, D. G. & Endsley, M. R. (1996). Sources of situation awareness errors in aviation. *Aviation, Space, & Environmental Medicine, 67*(6), 507–512.

Jones, R. K. (2003). Miscommunication between pilots and air traffic control. *Language Problems & Language Planning, 27*(3), 233–248.

Kanarish, J. (2017). *Listen for your call sign*. Retrieved from http://atccommunication.com/listen-for-your-call-sign#comments.

Keatley, C. W. (1992). History of bilingualism research in cognitive psychology. In R. J. Harris (Ed.), *Cognitive processing in bilinguals* (Vol. 4, pp.15–49). Amsterdam: Elsevier.

Kellogg, R. T. (2016). *Fundamentals of cognitive psychology*. Los Angeles: Sage.

Klein, G. (2000). Analysis of situation awareness from critical incident reports. In M. R. Endsley & D. J. Garland (Eds.), *Situation awareness analysis and measurement* (pp. 51–71). Mahwah, NJ: Laurence Erlbaum Associates.

Knold (2007, July 6). Re: CNN story on Chinese pilots and their English skills [Online forum comment]. Retrieved from https://www.pprune.org/rumours-news/283050-cnn-story-chinese-pilots-their-english-skills.html.

Kroll, J. F., Bobb, S. C., & Wodniecka, Z. (2006). Language selectivity is the exception, not the rule: Arguments against a fixed locus of language selection in bilingual speech. *Bilingualism: Language & Cognition, 9*(2), 119–135.

Krueger, L. E. (1986). Why 2 × 2 = 5 looks so wrong: On the odd–even rule in product verification. *Memory & Cognition*, *14*(2), 141–149.

Kuipers, J. R. & Thierry, G. (2010). Event-related brain potentials reveal the time-course of language change detection in early bilinguals. *NeuroImage*, *50*, 1633–1638.

Kunde, D. (2005). *Event prediction for modelling mental simulation in naturalistic decision making*. PhD thesis. Monterey, CA: Naval Postgraduate School.

Lamm, E. & Lawrence, N. (2010, July 12). *Interior sound levels in general aviation aircraft*. Retrieved from https://ohsonline.com/articles/2010/07/12/interior-sound-levels-in-general-aviation-aircraft.aspx?m=1.

Lehto, M. & Landry, S. J. (2013). *Introduction to human factors and ergonomics for engineers* (2nd ed.). Boca Raton, FL: CRC Press.

Lei, M., Akama, H., & Murphy, B. (2014). Neural basis of language switching in the brain: fMRI evidence from Korean–Chinese early bilinguals. *Brain & Language*, *138*, 12–18.

Liu, H., Fan, N., Rossi, S., Yao, P., & Chen, B. (2016). The effect of cognitive flexibility on task switching and language switching. *International Journal of Bilingualism*, *20*(5), 563–579.

Lloyd-Jones, G. & Binch, Ch. (2012). A case study evaluation of the English language progress of Chinese students on two UK postgraduate engineering courses. *IELTS Research Reports, 13*.

Lomov, B. F., Duškov, B. A., Rubachin, V. F., & Smirnov, B. A. (1983). *Základy inžinierskej psychológie [Fundamentals of engineering psychology]*. Bratislava: Slovenské Pedagogické Nakladateľstvo.

Los, S. A. (1996). On the origin of mixing costs: Exploring information processing in pure and mixed blocks of trials. *Acta Psychologica*, *94*, 145–188.

Los, S. A. (1999). Identifying stimuli of different perceptual categories in pure and mixed blocks of trials: Evidence for stimulus-driven switch costs. *Acta Psychologica*, *10*3, 173–205.

Lupker, S. J., Kinoshita, S., Coltheart, M., & Taylor, T. E. (2003). Mixing costs and mixing benefits in naming words, pictures, and sums. *Journal of Memory & Language*, *49,* 556–575.

Lynn, S. K. & Barrett, L.F. (2014). “Utilizing” signal detection theory. *Psychological Science*, *25*(9), 1663–1673.

MacDonald, J. A. & Balakrishnan, J. D. (2005). *Signal detection theory: From “Encyclopedia of cognitive science”.* Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.667.7678&rep=rep1&type=pdf.

Macizo, P., Bajo, T., & Paolieri, D. (2012). Language switching and language competition. *Second Language Research*, *28*(2), 131–149.

Marsh, L. & Maki, R. (1976). Efficiency of arithmetic operations in bilinguals as a function of language. *Memory & Cognition, 4,* 459–464.

Marslen-Wilson, W. D. (1985). Speech shadowing and speech comprehension. *Speech Communication*, *4,* 55–73.

Marslen-Wilson, W. D. & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, *8,* 1–71.

Martin, M. C., Macizo, P., & Bajo, T. (2010). Time course of inhibitory processes in bilingual language processing. *British Journal of Psychology*, *101*, 679–693.

Massey University, (2017). *Entry requirements for international students*. Retrieved from http://www.massey.ac.nz/massey/international/study-with-massey/entry-equirements/entry-requirements_home.cfm.

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language & Cognitive Processes*, *27*(7/8), 953–978.

McAnally, K. I., Martin, R. L., Eramudugolla, R., Stuart, G. W., Irvine, D. R. F., & Mattingley, J. B. (2010). A dual-process account of auditory change detection. *Journal of Experimental Psychology: Human Perception & Performance*, *36*(4), 994–1004.

McClain, L. & Huang, J. Y. S. (1982). Speed of simple arithmetic in bilinguals. *Memory & Cognition*, *10*(6), 591–596.

McMillan, D. (1998). *"Say again?" Miscommunications in air traffic control*. Masters project. Australia: Queensland University of Technology. Retrieved from http://www.aero-lingo.com/docs/Miscommunications%20in%20
Air%20Traffic%20Control.pdf.

McNicol, D. (1972). *A primer of signal detection theory*. London: Allen and Uwin.

Mehrabian, A. (1981). *Silent messages: Implicit communication of emotions and attitudes*. Belmont, California: Wadsworth Pub. Co.

Merrifield, G. (2012). The use of IELTS for assessing immigration eligibility in Australia, New Zealand, Canada and the United Kingdom. *IELTS Research Reports, 13*.

Meuter, R. F. I. & Allport, A. (1999). Bilingual language switching in naming: Asymmetrical costs of language selection. *Journal of Memory & Language*, *40*(1), 25–40.

Michael Good Videos (2013, September 5). Poor English skills from Air China Pilot [Video file]. Retrieved from https://www.youtube.com/watch?v=hUdqyAIAHNQ.

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.

Miller, K., Major, S., Shu, H., & Zhang, H. (2000). Ordinal knowledge: Number names and number concepts in Chinese and English. *Canadian Journal of Experimental Psychology*, *54*(2), 129–139.

Miller, K., Smith, C., Zhu, J., & Zhang, H. (1995). Preschool origins of cross-national differences in mathematical competence: The role of number-naming systems. *Psychological Science*, *6*, 56–60.

Miller, K. & Stigler, J. (1987). Computing in Chinese: Cultural variation in a basic cognitive skill. *Cognitive Development*, *2*, 279–305.

Molesworth, B., R. C. & Estival, D. (2015). Miscommunication in general aviation: the influence of external factors on communication errors. *Safety Science*, *73*, 73–79.

Monan, W. P. (1988). *Human factors in aviation operations: The hearback problem*. National Aeronautics and Space Administration Contractor Report 177398. Moffatt Field, CA: Ames Research Center.

Monan, W. P. (1991). Readback/hearback. *ASRS Directline Newsletter by the Analyst of NASA's Aviation Safety Reporting System*. Retrieved from https://asrs.arc.nasa.gov/ publications/directline/dl1_read.htm.

Moradi, H. (2014). An investigation through different types of bilinguals and bilingualism. *International Journal of Humanities & Social Science Studies*, *1*(2), 107–112.

Moray, N. (1959). Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology*, 56–60.

Moreno, S., Bialystok, E., Wodniecka, Z., & Alain, C. (2010). Conflict resolution in sentence processing by bilinguals. *Journal of Neurolinguistics, 23*, 564–79.

Moreno, E. M., Federmeier, K. D., & Kutas, M. (2002). Switching languages, switching palabras (words): An electrophysiological study of code switching. *Brain & Language*, *80*, 188–207.

Mori, D. (2010, March 22). Air traffic control: Swiss Airbus bird strike [Video file]. Retrieved from https://www.youtube.com/watch?v=lICb8p9SvvM.

Murray, I. R., Baber, Ch., & South, A. (1996). Towards a definition and working model of stress and its effects on speech. *Speech Communication*, 3–12.

National Transportation Safety Board. (1990). *Aircraft accident report.* Avianca, the airline of Columbia, Boeing 707-321B, HK 2016. Fuel exhaustion. Cove Neck, New York, January 25, 1990. Washington, D.C. 20594. NTSB/AAR-91/04.

National Transportation Safety Board. (1973). *Aircraft accident report*. Eastern Air Lines, Inc. L-1011, N310EA. Miami, Florida, December 29, 1972. Washington, D.C. 20591. NTSB-AAR-73-14.

Neisser, U. & Becklen, R. (1975). Selective looking: Attending to visually specified events. *Cognitive Psychology*, *7*, 480–494.

Nevile, M. (2004). *Beyond the black-box. Talk-in-interaction in the airline cockpit*. Aldershot, UK: Ashgate Publishing.

Nordquist, R. (2017, May 28). *What is a language family?* Retrieved from https://www.thoughtco.com/what-is-a-language-family-1691216.

Oberg, B. (2016). *The difference between the Chinese and English languages*. Retrieved from https://www.linkedin.com/pulse/difference-between-chinese-english-languages-bill-oberg.

Orasanu, J., Fischer, U. & Davison, J. (1997). Cross-cultural barriers to effective communication in aviation, In S. Oskamp & C. Granrose (Eds.), *Cross-cultural work groups: The Claremont Symposium on Applied Social Psychology* (pp. 1–23). Sage Publications.

Orlady, H. W. & Orlady, L. M. (1999*). Human factors in multi-crew flight operations*. Aldershot, UK: Ashgate.

Pahlow, H. (n.d.). *Table of English tenses*. Retrieved from https://www.ego4u.com/en/cram-up/grammar/tenses.

Palacios, R., Doshi, A., & Gupta, A. (2008). Computing aircraft position prediction. *Open Transportation Journal*, *2*, 94–97.

Pannese, A., Herrmann, Ch. S., & Sussman, E. (2015). Analysing the auditory scene: Neurophysiologic evidence of a dissociation between detection of regularity and detection of change. *Brain Topography*, *28*(3), 411–422.

Patrick, J. & James, N. (2004). A task-oriented perspective of situation awareness. In S. Banbury & S. Tremblay (Eds.), *A cognitive approach to situation awareness: Theory and application* (pp. 61–81). Burlington, VT: Ashgate.

Pavlenko, A. (2014). *The bilingual mind: and what it tells us about language and thought*. Cambridge University Press.

Peirce, J. W. (2007). PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods, 162*(1–2), 8–13.

Peterson, L. R. & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology, 58*(3), 193–198.

Pew, R. W. (2000). The state of situation awareness measurement: Heading toward the next century, In M. R. Endsley & D. J. Garland (Ed.), *Situation awareness analysis and measurement* (pp. 33–47). Mahwah, NJ: Laurence Erlbaum Associates.

Philipp, A. M., Gade, M., & Koch, I. (2007). Inhibitory processes in language switching: Evidence from switching language defined response sets. *European Journal of Cognitive Psychology, 19,* 395–416.

Philipp, A. M., Kalinich, C., Koch, I., & Schubotz, R. I. (2008). Mixing costs and switch costs when switching stimulus dimensions in serial predictions. *Psychological Research, 72,* 405–414.

Philipp, A. M. & Koch, I. (2011, September). *The role of lexical selection and speech production in language switching*. Poster presented at the 17th Meeting of the European Society for Cognitive Psychology, San Sebastian, Spain.

Piasecka, L. (2013). What does it feel like to use English? Empirical evidence from EFL students. In E. Piechurska-Kuciel & E. Szymanska-Czaplak (Eds.), *Language in cognition and affect, second language learning and teaching* (pp. 219–237). Heidelberg: Springer-Verlag.

Prinzo, O. V. & Campbell, A. (2008). *U.S. airline transport pilot international flight language experiences, Report 1: Background information and general/pre-flight preparation*. Washington: FAA Civil Aerospace Medical Institute.

Prinzo, O. V., Campbell, A., Hendrix, A., & Hendrix, R. (2010a). *U.S. airline transport pilot international flight language experiences, Report 3: Language experiences in non-native English-speaking airspace/airports*. Washington: FAA Civil Aerospace Medical Institute.

Prinzo, O. V., Campbell, A., Hendrix, A., & Hendrix, R. (2010b). *U.S. airline transport pilot international flight language experiences, Report 4: Non-native English-speaking* *controllers communicating with native English-speaking pilots*. Washington. FAA Civil Aerospace Medical Institute.

Prinzo, O. V., Campbell, A., Hendrix, A., & Hendrix, R. (2011). *U.S. airline transport pilot international flight language experiences, Report 6: Native English-speaking controllers communicating with non-native English-speaking pilots*. Washington. FAA Civil Aerospace Medical Institute.

Prinzo, O. V., Hendrix, A. M., & Hendrix, R. (2006). *The outcome of ATC message complexity on pilot readback performance*. Washington: FAA Civil Aerospace Medical Institute. Final report DOT/FAA/AM-06/25.

Prinzo, O. V. & Morrow, D. G. (2002). Improving pilot/air traffic control voice communication in general aviation. *International Journal of Aviation Psychology*. *12*(4), 341–357.

Prinzo, O. V. & Thompson, A. C. (2009). *The ICAO English Language Proficiency Rating Scale applied to enroute voice communications of U.S. and foreign pilots*. Washington: FAA Civil Aerospace Medical Institute.

Prior, A. & Gollan, T. H. (2013). The elusive link between language control and executive control: A case of limited transfer. *Journal of Cognitive Psychology*, *25*(5), 622–645.

Rayner, J. & Ellis, A.W. (2007). The control of bilingual language switching. In A. S. Meyer, L. R. Wheeldon, & A. Krott. (Eds.), *Automaticity and control in language processing* (pp. 43–62). New York: Psychology Press.

Read, J., Wette, R., & Deverall, P. (2009). Achieving English proficiency for professional registration: The experience of overseas-qualified health professionals in the New Zealand context. *IELTS Research Reports, 10*.

Reder, L. M. (1982). Plausibility judgments versus fact retrieval: Alternative strategies for sentence verification. *Psychological Review, 89*(3), 248–278.

Rediff on the net (n.d.). *Communication gap caused Charkhi Dadri mishap: ATC guild*. Retrieved from http://www.rediff.com/news/may/15aai.htm.

Riley, J. M., Endsley, M. E., Bolstad, C. H., & Cuevas, H. M. (2006). Collaborative planning and situation awareness in Army command and control. *Ergonomics, 49*(12–13); 1139–1153.

Rodriguez-Fornells, A., van der Lugt, A., Rotte, M., Britti, B., Heinze, H. J., & Munte, T. F. (2005). Second language interferes with word production in fluent bilinguals: Brain potential and functional imaging evidence. *Journal of Cognitive Neuroscience, 17*(3), 422–433.

Rosinski, D. J. (2010). Sterile cockpit or not: It's all about team and effective communication. *Journal of Thoracic and Cardiovascular Surgery, 140*, 10–11.

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology, 1*(1), 43–46.

Rousseau, R., Tremblay, S., & Breton, R. (2004). Defining and modelling situation awareness: A critical review. In S. Banbury & S. Tremblay (Eds.), *A cognitive approach to situation awareness: Theory and application* (pp. 3–21). Burlington, VT: Ashgate.

Ruigendijk, E., Zeller, J. P., & Hentschel, G. (2009, October). *How L2-learners' brains react to codeswitches: An ERP study*. Paper presented at the University of Amsterdam.

Sarter, N. B. & Woods, D. D. (1991). Situation awareness: a critical but ill-defined phenomenon. *International Journal of Aviation Psychology, 1*, 45–57.

Sarter, N. B. & Woods, D. D. (1995). How in the world did we ever get into that mode? Mode error and awareness in supervisory control. *Human Factors*, *37*(1), 5–19.

Saslow, L. R., McCoy, S., van der Lowe, I., Cosley, B., Vartan, A., Oveis, Ch., & Epel, E. S. (2014). Speaking under pressure: Low linguistic complexity is linked to high physiological and emotional stress reactivity. *Psychophysiology*, *51*, 257–266.

Schlimm, F. (2017, May 31). *Can Chinese be airline pilots?* Retrieved from https://www.quora.com/Can-Chinese-be-airline-pilots.

Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology*, *6*(4), 147–151.

Schmidt, C. O. & Kohlmann, T. (2008). When to use the odds ratio or the relative risk? *International Journal of Public Health*, *53*, 165–167.

Seltman, H. J. (2015). *Experimental design and analysis*. Retrieved from http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf.

Servan-Schreiber, D. (2004). *The instinct to heal: curing depression, anxiety and stress without drugs and without talk therapy*. Emmaus, PA: Rodale Press.

Shebilske, W. L., Goettl, B. P., & Garland, D. J. (2000). Situation awareness, automaticity, and training. In M. R. Endsley & D. J. Garland (Eds.), *Situation awareness analysis and measurement* (pp. 303–323). Mahwah, NJ: L. Erlbaum Associates.

Shepard, R. N., Kilpatric, D. W., & Cunningham, J. P. (1975). The internal representation of numbers. *Cognitive Psychology*, *7*, 8–138.

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences*, *12*(5), 182–186.

Simmon, D. A. (1998). Boeing 757 CFIT Accident at Cali, Colombia, becomes focus of lessons learned. *Flight Safety Digest*, *17*(5/6), 1–31.

Simpson, B. D., Bolia, R. S., McKinley, R. L., & Brungart, D. S. (2005). *The impact of hearing protection on sound localization and orienting behaviour*. Air Force Research laboratory. Springfield, Virginia.

SKYbrary (2013, January 29). *Frequency congestion*. Retrieved from https://www.skybrary.aero/index.php/Frequency_Congestion.

Smith, H. & Haslett, S. (2007). Attitudes of tertiary key decision-makers towards English language tests in Aotearoa New Zealand: Report on the results of a national provider survey. *IELTS Research Reports, 7*.

Smith, H. & Haslett, S. (2008). Use of the IELTS General Training module in technical and vocational tertiary institutions: A case study from Aotearoa New Zealand. *IELTS Research Reports, 8*.

Smith, P. G., Morrow, R. H., & Ross, D. A. (2015). *Field trials of health interventions: A toolbox*. Oxford University Press.

Smith, M. S. & Stein, M. K. (1998). Selecting and creating mathematical tasks: From research to practice. *Mathematics Teaching in the Middle School, 3*(5), 344–350.

Spelke, E. S. & Tsivkin, S. (2001). Language and number: a bilingual training study. *Cognition, 78*(1), 45–88.

Stanislaw, H. & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*(1), 137–149.

Sternberg, S. (1969). Memory-scanning: Mental processes revealed by reaction-time experiments. *American Scientist, 57*(4), 421–457.

Steve (2013, January 30). Re: Gain vs. Amplify vs. Compressor [Online forum comment]. Retrieved from http://forum.audacityteam.org/viewtopic.php?f=16&t=70729.

Stigler, J., Lee, S., & Stevenson, H. (1986). Digit memory in Chinese and English: Evidence for a temporally limited store. *Cognition, 23*, 1–20.

Southwest EcoMotoring Club (n.d.). *Predict the road ahead (Anticipation*). Retrieved from https://southwestecomotoring.com/eco-drivers-ed/predict-the-road-ahead-anticipation/.

Steve (2018, March 15). Re: Convert negative dB scaling to positive [Online forum comment]. Retrieved from https://forum.audacityteam.org/viewtopic.php?f=46&t=99271.

Sulistyawati, K., Wickens, C. D., & Chui, Y. P. (2011). Prediction in situation awareness: confidence bias and underlying cognitive abilities. *International Journal of Aviation Psychology, 2*(2), 153–174.

Sumwalt, R. L. (1993). The sterile cockpit. *ASRS Directline, 4*, 18–22. Retrieved from https://asrs.arc.nasa.gov/publications/directline/dl4_sterile.htm.

Sumwalt, R. L. & Watson, A. W. (1995). ASRS incident data reveal details of flight-crew performance during aircraft malfunctions. *Flight Safety Digest, 14*(10), 1–7.

Suresh, K. P. (2011). An overview of randomization techniques: An unbiased assessment of outcome in clinical research. *Journal of Human Reproductive Sciences, 4*(1), 8–11.

SurveyMonkey (n.d.). *Using skip logic in a Survey*. Retrieved from https://www.surveymonkey.com/mp/tour/skiplogic/.

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics*. Collected Papers. New Jersey: Lawrence Erlbaum Associates.

Swets, J. A. (2001). Signal detection theory, history of. In N. J. Smelser, & P. B. Baltes (Eds.), *International Encyclopedia of the Social & Behavioral Sciences* [e-book], 14078–14082. https://doi.org/10.1016/B0-08-043076-7/00678-1

Swets, J. A., Tanner, W. P., Jr., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review*, *68*(5), 301–340.

Symbiotics Ltd. (n.d.). *Full ADAPT Intelligent Selection for pilots*. Retrieved online from https://www.symbioticsltd.co.uk/assessment-selection/full-adapt/.

Tabachnick, B. G. & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Needham Heights, MA: Pearson Education.

Tajima, A. (2004). Fatal miscommunication: English in aviation safety. *World Englishes, 23*(3), 451–470.

Tamamaki, K. (1993). Language dominance in bilinguals' arithmetic operations according to their language use. *Language Learning*, *43*(2), 239–262.

Tao, L., Marzeocová, A., Taft, M., Asanowicz, D., & Wodniecka, Z. (2011). The efficiency of attentional networks in early and late bilinguals: The role of age of acquisition. *Frontiers in Psychology*, *2*(123), 1–19.

Tarlowski, A., Wodniecka, Z., & Marzecová, A. (2012). Language switching in the production of phrases. *Journal of Psycholinguistic Research*, *42*, 103–118.

Terenzi, M., Iyer, N., Simpson, B. D., Bolia, R. S., & Di Nocera, F. (2006). Using spatial intercoms to improve speech intelligibility for international teams. *Proceedings of the Human Factors and Ergonomics Society,* 50th Annual Meeting.

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

Vaid, J. & Menon, R. (2000). Correlates of bilinguals' preferred language for mental computations. *Spanish Applied Linguistics, 4(2), 325–342.*

VandenBos, G. R. (2015). *APA dictionary of psychology* (2nd ed.). Washington DC: American Psychological Association.

Vandergrift, L. & Baker, S. (2015). Learner variables in second language listening comprehension: An exploratory path analysis. *Language Learning, 65*(2), 390–416.

Van Rinsveld, A., Schiltz, Ch., Brunner, M., Landerl, K., & Ugen, S. (2016). Solving arithmetic problems in first and second language: Does the language context matter? *Learning & Instruction*, *42*, 72–82.

Venkatraman, V., Siong, S. Ch., Chee, M. W. L., & Ansari, D. (2006). Effect of language switching on arithmetic: A bilingual fMRI study. *Journal of Cognitive Neuroscience*, *18*(1), 64–74.

Verhoef, K., Roelofs, A., & Chwilla, D. J. (2009). Role of inhibition in language switching: Evidence from event-related brain potentials in overt picture naming. *Cognition*, *110*, 84–99.

Viera, A. J. & Bangdiwala, S. I. (2007). Eliminating bias in randomized controlled trials: importance of allocation concealment and masking. *Family Medicine, 39*(2), 132–137.

Wadhera, R. K., Parker, S. H., Burkhart, H. M., Greason, K. L., Neal, J. R., Levenick, K. M., Wiegmann, D. A., & Sundt, T. M. (2010). Is the "sterile cockpit" concept applicable to cardiovascular surgery critical intervals or critical events? The impact of protocol-driven communication during cardiopulmonary bypass. *Journal of Thoracic & Cardiovascular Surgery, 139*, 312–319.

Wang, Y. & Chen, J. (2013). Differences of English and Chinese as written languages and strategies in English writing teaching. *Theory & Practice in Language Studies*, *3*(4), 647–652.

Wang, Y., Lin, L., & Hirsch, J. (2007). Mathematical and linguistic processing differs between native and second languages: An fMRI study. *Brain Imaging & Behavior*, *1*, 68–82.

Weissberger, G. H., Gollan, T. H., Bondi, M. W., Clark, L. R., & Wierenga, Ch. E. (2015). Language and task switching in the bilingual brain: Bilinguals are staying, not switching, experts. *Neuropsychologia, 66*, 193–203.

Widaman, K. E, Geary, D. C., Cormier, P., & Little, T. D. (1989). A componential model for mental addition. *Journal of Experimental Psychology, Learning, Memory, & Cognition, 15*, 898–919.

Willan, P. (2001). 118 killed as jet crashes at Milan airport. *The Guardian*. Retrieved from https://www.theguardian.com/world/2001/oct/09/philipwillan.

Winke, P. & Lim, H. (2014). The effects of testwiseness and test-taking anxiety on L2 listening test performance: A visual (eye-tracking) and attentional investigation. *IELTS Research Report, 3*.

Wintre, G. M., North, C., & Sugar, L. A. (2001). Psychologists' response to criticisms about research based on undergraduate participants: A developmental perspective. *Canadian Psychology/Psychologie Canadienne, 42*(3), 216–225.

Wray, A. & Pegg, Ch. (2009). The effect of memorized learning on the writing scores of Chinese IELTS test-takers. *IELTS Research Reports, 9*.

Wu, X., Yang, Z., Huang, Y., Chen, J., Li, L., Daneman, M., & Schneider, B. A. (2011). Cross-language differences in informational masking of speech by speech: English versus Mandarin Chinese. *Journal of Speech Language & Hearing Research, 54*, 1506–1524.

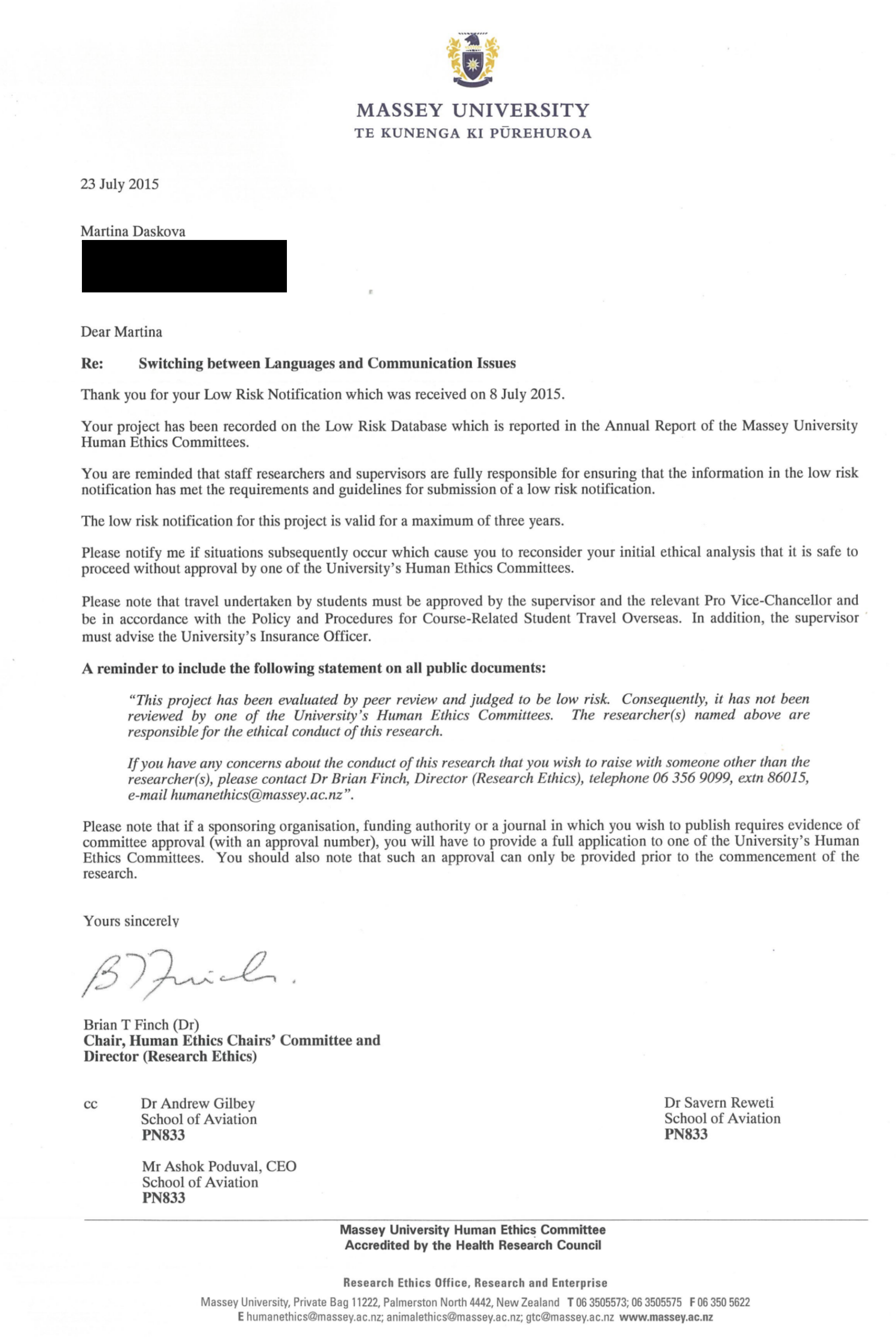
Xinhua News Agency (2007, June 23). Tough English test could ground Chinese pilots. *CHINA.ORG.CN*. Retrieved from http://www.china.org.cn/english/China/214901.htm.

Ye, Z. & Zhou, X. (2009). Executive control in language processing. *Neuroscience & Behavioral Reviews, 33*, 1168–1177.

Zheng, X., Roelofs, A., & Lemhofer, K. (2018). Language selection errors in switching: Language priming or cognitive control? *Language, Cognition & Neuroscience, 33*(2), 139–147.

Zwitserlood, P. (1998). Spoken words in sentence contexts. In A. D. Friederici (Ed.) *Language comprehension: A biological perspective*, (pp. 71–99). Heidelberg: Springer-Verlag Berlin.

APPENDIX A: LOW-RISK ETHICS NOTIFICATION FOR STUDY 1



MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA

23 July 2015

Martina Daskova
[redacted]

Dear Martina

Re: Switching between Languages and Communication Issues

Thank you for your Low Risk Notification which was received on 8 July 2015.

Your project has been recorded on the Low Risk Database which is reported in the Annual Report of the Massey University Human Ethics Committees.

You are reminded that staff researchers and supervisors are fully responsible for ensuring that the information in the low risk notification has met the requirements and guidelines for submission of a low risk notification.

The low risk notification for this project is valid for a maximum of three years.

Please notify me if situations subsequently occur which cause you to reconsider your initial ethical analysis that it is safe to proceed without approval by one of the University's Human Ethics Committees.

Please note that travel undertaken by students must be approved by the supervisor and the relevant Pro Vice-Chancellor and be in accordance with the Policy and Procedures for Course-Related Student Travel Overseas. In addition, the supervisor must advise the University's Insurance Officer.

A reminder to include the following statement on all public documents:

> *"This project has been evaluated by peer review and judged to be low risk. Consequently, it has not been reviewed by one of the University's Human Ethics Committees. The researcher(s) named above are responsible for the ethical conduct of this research.*
>
> *If you have any concerns about the conduct of this research that you wish to raise with someone other than the researcher(s), please contact Dr Brian Finch, Director (Research Ethics), telephone 06 356 9099, extn 86015, e-mail humanethics@massey.ac.nz".*

Please note that if a sponsoring organisation, funding authority or a journal in which you wish to publish requires evidence of committee approval (with an approval number), you will have to provide a full application to one of the University's Human Ethics Committees. You should also note that such an approval can only be provided prior to the commencement of the research.

Yours sincerely

Brian T Finch (Dr)
Chair, Human Ethics Chairs' Committee and
Director (Research Ethics)

cc Dr Andrew Gilbey
School of Aviation
PN833

Dr Savern Reweti
School of Aviation
PN833

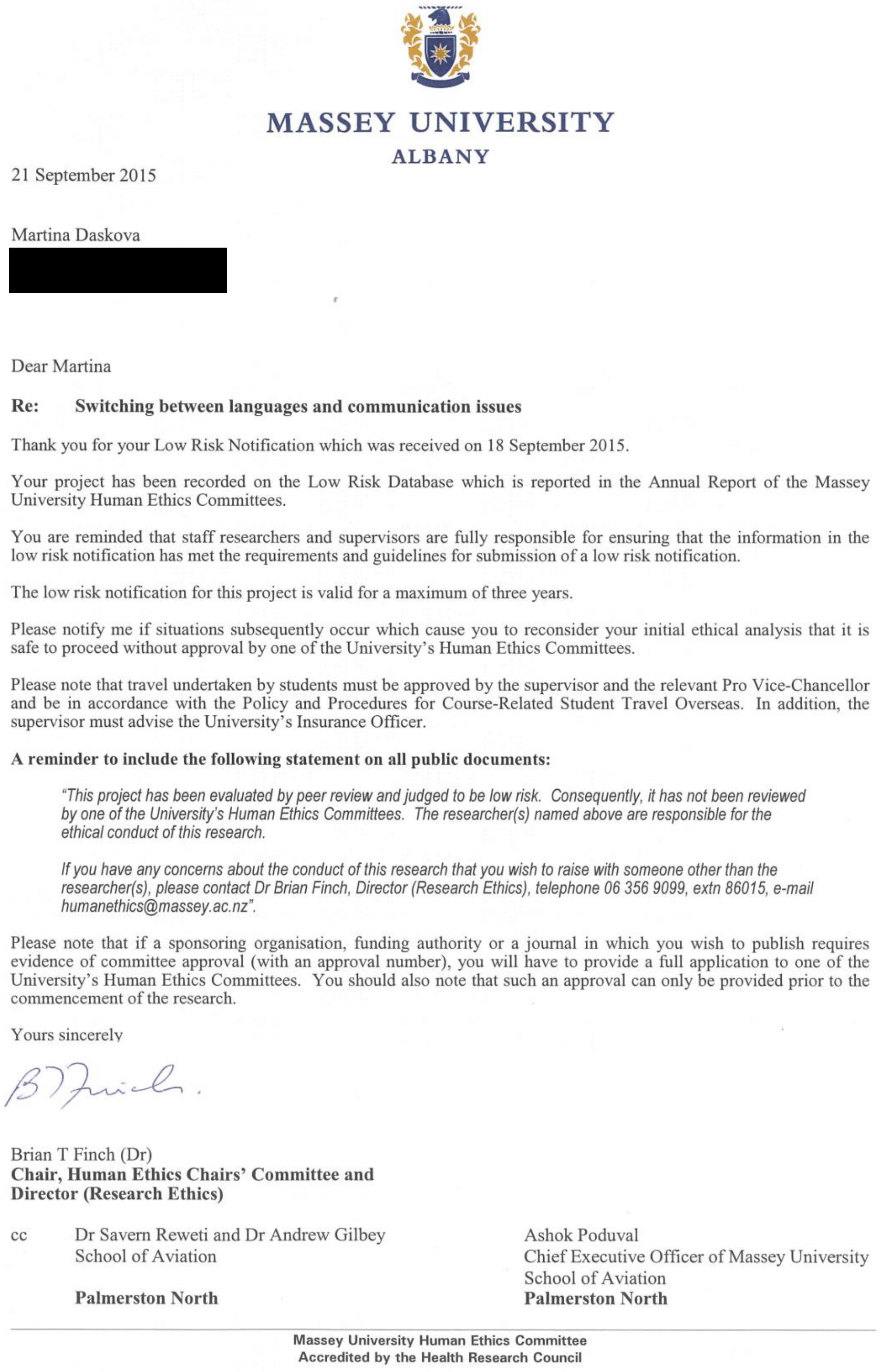
Mr Ashok Poduval, CEO
School of Aviation
PN833

Massey University Human Ethics Committee
Accredited by the Health Research Council

Research Ethics Office, Research and Enterprise

Massey University, Private Bag 11222, Palmerston North 4442, New Zealand **T** 06 3505573; 06 3505575 **F** 06 350 5622
E humanethics@massey.ac.nz; animalethics@massey.ac.nz; gtc@massey.ac.nz **www.massey.ac.nz**

APPENDIX B: LOW-RISK ETHICS NOTIFICATION FOR STUDY 2



MASSEY UNIVERSITY
ALBANY

21 September 2015

Martina Daskova
[redacted]

Dear Martina

Re: Switching between languages and communication issues

Thank you for your Low Risk Notification which was received on 18 September 2015.

Your project has been recorded on the Low Risk Database which is reported in the Annual Report of the Massey University Human Ethics Committees.

You are reminded that staff researchers and supervisors are fully responsible for ensuring that the information in the low risk notification has met the requirements and guidelines for submission of a low risk notification.

The low risk notification for this project is valid for a maximum of three years.

Please notify me if situations subsequently occur which cause you to reconsider your initial ethical analysis that it is safe to proceed without approval by one of the University's Human Ethics Committees.

Please note that travel undertaken by students must be approved by the supervisor and the relevant Pro Vice-Chancellor and be in accordance with the Policy and Procedures for Course-Related Student Travel Overseas. In addition, the supervisor must advise the University's Insurance Officer.

A reminder to include the following statement on all public documents:

> *"This project has been evaluated by peer review and judged to be low risk. Consequently, it has not been reviewed by one of the University's Human Ethics Committees. The researcher(s) named above are responsible for the ethical conduct of this research.*
>
> *If you have any concerns about the conduct of this research that you wish to raise with someone other than the researcher(s), please contact Dr Brian Finch, Director (Research Ethics), telephone 06 356 9099, extn 86015, e-mail humanethics@massey.ac.nz".*

Please note that if a sponsoring organisation, funding authority or a journal in which you wish to publish requires evidence of committee approval (with an approval number), you will have to provide a full application to one of the University's Human Ethics Committees. You should also note that such an approval can only be provided prior to the commencement of the research.

Yours sincerely

Brian T Finch (Dr)
Chair, Human Ethics Chairs' Committee and
Director (Research Ethics)

cc Dr Savern Reweti and Dr Andrew Gilbey
School of Aviation

Palmerston North

Ashok Poduval
Chief Executive Officer of Massey University
School of Aviation
Palmerston North

Massey University Human Ethics Committee
Accredited by the Health Research Council

APPENDIX C: STIMULI LIST FOR STUDY 2

L1 Condition	*L2 Condition*	*Mix Condition*
Target	*Target*	*Target*
729	531	462
Similarity 2	*Similarity 2*	*Similarity 2*
720	537	463 Chinese
721	534	461 Chinese
725	532	467 Chinese
723	536	469 Chinese
724	535	464 English
726	538	460 English
728	539	468 English
727	530	465 English
Similarity 1	*Similarity 1*	*Similarity 1*
761	549	426 Chinese
713	594	493 Chinese
734	543	434 Chinese
738	526	456 Chinese
746	564	428 English
789	567	412 English
784	572	473 English
792	584	427 English
Similarity 0	*Similarity 0*	*Similarity 0*
862	961	786 Chinese
917	472	819 Chinese
825	379	256 Chinese
876	683	756 Chinese
137	142	298 Chinese
654	685	924 Chinese
987	629	548 Chinese
914	649	985 Chinese
324	384	245 Chinese
235	643	867 English
436	631	613 English
382	193	751 English
562	625	378 English
967	839	357 English
147	965	392 English
179	346	375 English
286	128	591 English
368	523	524 English

APPENDIX D: LOW-RISK ETHICS NOTIFICATION FOR STUDY 3

Human Ethics Notification – 4000016381

humanethics@massey.ac.nz <humanethics@massey.ac.nz> Jul 5, 2016 at 12:13 PM
To: A.Lindsay@massey.ac.nz, Martina.Daskova.1@uni.massey.ac.nz, A.P.Gilbey@massey.ac.nz
Cc: M.E.Thomas@massey.ac.nz

HoU Review Group

Ethics Notification Number: 4000016381
Title: Message comprehension in a language switching task

Thank you for your notification which you have assessed as Low Risk.

Your project has been recorded in our system which is reported in the Annual Report of the Massey University Human Ethics Committee.

The low risk notification for this project is valid for a maximum of three years.

If situations subsequently occur which cause you to reconsider your ethical analysis, please log on to http://rims.massey.ac.nz and register the changes in order that they be assessed as safe to proceed.

Please note that travel undertaken by students must be approved by the supervisor and the relevant Pro Vice-Chancellor and be in accordance with the Policy and Procedures for Course-Related Student Travel Overseas. In addition, the supervisor must advise the University's Insurance Officer.

A reminder to include the following statement on all public documents:

"This project has been evaluated by peer review and judged to be low risk. Consequently it has not been reviewed by one of the University's Human Ethics Committees. The researcher(s) named in this document are responsible for the ethical conduct of this research.
If you have any concerns about the conduct of this research that you want to raise with someone other than the researcher(s), please contact Dr Brian Finch, Director (Research Ethics), email humanethics@massey.ac.nz. "

Please note that if a sponsoring organisation, funding authority or a journal in which you wish to publish require evidence of committee approval (with an approval number), you will have to complete the application form again answering yes to the publication question to provide more information to go before one of the University's Human Ethics Committees. You should also note that such an approval can only be provided prior to the commencement of the research.

You are reminded that staff researchers and supervisors are fully responsible for ensuring that the information in the low risk notification has met the requirements and guidelines for submission of a low risk notification.

If you wish to print an official copy of this letter, please login to the RIMS system, and under the Reporting section, View Reports you will find a link to run the LR Report.

Yours sincerely

Dr Brian Finch
Chair, Human Ethics Chairs' Committee and
Director (Research Ethics)

APPENDIX E: STIMULI LIST FOR STUDY 3

L1 Condition

Correct

2 + 4 = 6
3 + 2 = 5
4 + 6 = 10
6 + 3 = 9
7 + 4 = 11
8 + 6 = 14
9 + 5 = 14
10 + 7 = 17
12 + 4 = 16
13 + 5 = 18
14 + 6 = 20
15 + 2 = 17
16 + 3 = 19
19 - 15 = 4
18 - 2 = 16
17 - 15 = 2
16 - 13 = 3
14 - 5 = 9
13 - 12 = 1
12 - 10 = 2
11 - 9 = 2
9 - 7 = 2
8 - 5 = 3
7 - 3 = 4
6 - 1 = 5
4 - 2 = 2

Incorrect

1 + 7 = 10
2 + 3 = 8
3 + 5 = 10
5 + 9 = 13
6 + 5 = 1
7 + 6 = 9
8 + 1 = 7
9 - 2 = 1
8 - 1 = 6
7 - 5 = 3
6 - 0 = 5
5 - 0 = 8
4 - 3 = 7
3 - 1 = 4

L2 Condition

Correct

2 + 9 = 11
4 + 5 = 9
5 + 2 = 7
6 + 4 = 10
7 + 10 = 17
8 + 4 = 12
11 + 3 = 14
12 + 2 = 14
13 + 2 = 15
14 + 1 = 15
15 + 3 = 18
16 + 4 = 20
17 + 2 = 19
19 - 10 = 9
18 - 17 = 1
17 - 13 = 4
16 - 9 = 7
13 - 11 = 2
11 - 8 = 3
10 - 7 = 3
9 - 4 = 5
8 - 7 = 1
7 - 2 = 5
6 - 2 = 4
5 - 3 = 2
4 - 1 = 3

Incorrect

2 + 7 = 8
3 + 4 = 5
4 + 9 = 6
5 + 3 = 7
6 + 2 = 9
7 + 3 = 4
8 + 3 = 5
10 – 8 = 7
9 – 5 = 6
8 – 6 = 4
7 – 6 = 2
5 – 6 = 11
4 – 4 = 8
3 – 2 = 0

MIX Condition

Correct

3 + 11 = 14 Chinese
5 + 7 = 12 Chinese
7 + 8 = 15 Chinese
9 + 3 = 12 Chinese
11 + 8 = 19 Chinese
15 + 4 = 19 Chinese
17 + 3 = 20 Chinese
2 + 12 = 14 Chinese
4 + 7 = 11 English
6 + 7 = 13 English
8 + 9 = 17 English
12 + 8 = 20 English
14 + 3 = 17 English
16 + 2 = 18 English
18 + 1 = 19 English
1 + 11 = 12 English
20 - 11 = 9 Chinese
16 - 14 = 2 Chinese
14 - 3 = 11 Chinese
12 - 5 = 7 Chinese
8 - 2 = 6 Chinese
6 - 5 = 1 Chinese
4 - 0 = 4 Chinese
20 - 6 = 14 Chinese
19 - 18 = 1 English
17 - 14 = 3 English
13 - 10 = 3 English
11 - 4 = 7 English
9 - 6 = 3 English
5 - 2 = 3 English
3 - 0 = 3 English
19 - 7 = 12 English

Incorrect

1 + 8 = 10 Chinese
3 + 6 = 7 Chinese
5 + 4 = 3 Chinese
7 + 1 = 6 Chinese
9 + 6 = 18 Chinese
2 + 5 = 10 English
6 + 1 = 5 English
8 + 2 = 12 English
10 + 4 = 20 English
10 - 2 = 12 Chinese
8 - 4 = 1 Chinese
6 - 3 = 4 Chinese
2 - 2 = 3 Chinese
9 - 1 = 10 English
7 - 4 = 2 English
5 - 1 = 2 English
3 - 3 = 1 English
1 - 1 = 3 English

APPENDIX F: LOW-RISK ETHICS NOTIFICATION FOR STUDY 4

Human Ethics Notification - 4000017320

humanethics@massey.ac.nz <humanethics@massey.ac.nz> Mar 20, 2017 at 9:22 AM
To: A.Lindsay@massey.ac.nz, Martina.Daskova.1@uni.massey.ac.nz, A.P.Gilbey@massey.ac.nz
Cc: M.E.Thomas@massey.ac.nz

HoU Review Group

Ethics Notification Number: 4000017320
Title: Prediction under different language conditions

Thank you for your notification which you have assessed as Low Risk.
Your project has been recorded in our system which is reported in the Annual Report of the Massey University Human Ethics Committee.

The low risk notification for this project is valid for a maximum of three years.

If situations subsequently occur which cause you to reconsider your ethical analysis, please log on to http://rims.massey.ac.nz and register the changes in order that they be assessed as safe to proceed.

Please note that travel undertaken by students must be approved by the supervisor and the relevant Pro Vice-Chancellor and be in accordance with the Policy and Procedures for Course-Related Student Travel Overseas. In addition, the supervisor must advise the University's Insurance Officer.

A reminder to include the following statement on all public documents:

"This project has been evaluated by peer review and judged to be low risk. Consequently it has not been reviewed by one of the University's Human Ethics Committees. The researcher(s) named in this document are responsible for the ethical conduct of this research.
If you have any concerns about the conduct of this research that you want to raise with someone other than the researcher(s), please contact Dr Brian Finch, Director (Research Ethics), email humanethics@massey.ac.nz. "

Please note that if a sponsoring organisation, funding authority or a journal in which you wish to publish require evidence of committee approval (with an approval number), you will have to complete the application form again answering yes to the publication question to provide more information to go before one of the University's Human Ethics Committees. You should also note that such an approval can only be provided prior to the commencement of the research.

You are reminded that staff researchers and supervisors are fully responsible for ensuring that the information in the low risk notification has met the requirements and guidelines for submission of a low risk notification.

If you wish to print an official copy of this letter, please login to the RIMS system, and under the Reporting section, View Reports you will find a link to run the LR Report.

Yours sincerely
Dr Brian Finch
Chair, Human Ethics Chairs' Committee and
Director (Research Ethics)

APPENDIX G: STIMULI LIST FOR STUDY 4

L1 Condition

Predicted Number (in brackets) was Present

1 2 4 5 7… (八; i.e., 8)
4 6 7 9 10… (十二; i.e., 12)
5 6 8 9 11… (十四; i.e., 14)
7 9 11 13 15… (二十一; i.e., 21)
8 11 14 17 20… (二十六; i.e., 26)
9 11 13 15 17… (二十五; i.e., 25)
10 12 13 15 16… (二十一; i.e., 21)
11 13 14 16 17… (二十三; i.e., 23)

Predicted Number (in brackets) was Not Present

2 4 6 8 10… (十一; i.e., 11)
3 6 9 12 15… (十六; i.e., 16)
6 8 10 12 14… (十七; i.e., 17)
10 13 16 19 22… (二十九; i.e., 29)
11 14 17 20 23… (三十三; i.e., 33)
0 2 3 5 6… (十; i.e., 10)
9 10 12 13 15… (二十; i.e., 20)
0 1 3 4 6… (十一; i.e., 11)

L2 Condition

Predicted Number (in brackets) was Present

1 3 4 6 7… (9)
2 4 5 7 8… (13)
4 7 10 13 16… (19)
5 7 8 10 11… (17)
7 9 10 12 13… (16)
11 13 15 17 19… (23)
0 3 6 9 12… (21)
2 5 8 11 14… (26)

Predicted Number (in brackets) was Not Present

3 4 6 7 9… (11)
6 7 9 10 12… (17)
8 10 12 14 16… (17)
9 12 15 18 21… (26)
10 11 13 14 16… (21)
1 3 5 7 9… (14)
3 5 7 9 11… (18)
4 5 7 8 10… (12)

Mix Condition

Predicted Number (in brackets) was Present

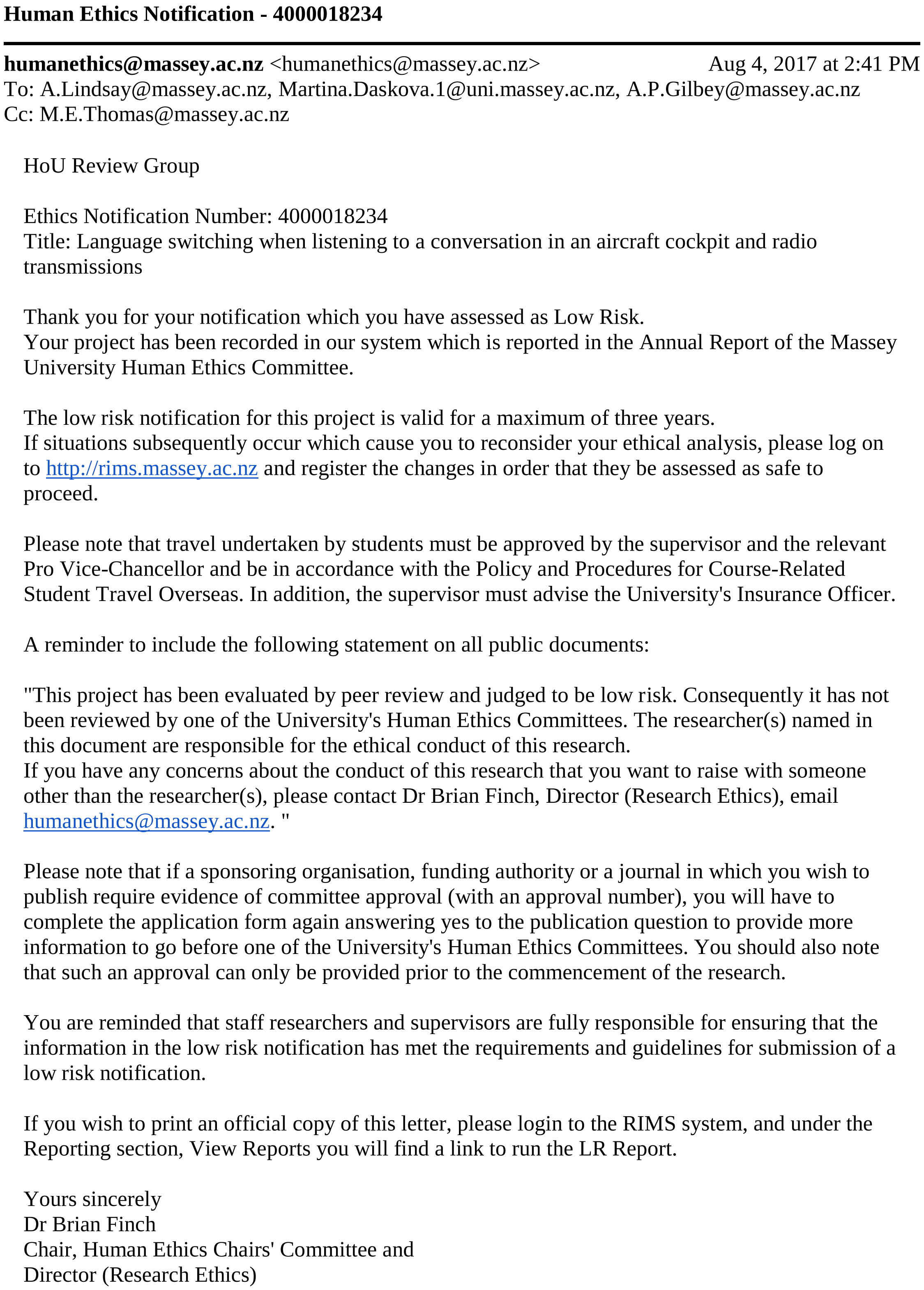
1Ch 4E 7Ch 10E 13Ch… (十六; i.e., 16)
3Ch 5Ch 6E 8E 9Ch… (11)
5E 8E 11Ch 14Ch 17E… (二十三; i.e., 23)
7Ch 10Ch 13E 16E 19Ch… (28)
8Ch 9E 11E 12E 14Ch… (十七; i.e., 17)
11Ch 12Ch 14E 15Ch 17E… (21)
5E 7Ch 9E 11Ch 13E… (21)
7Ch 8E 10E 11E 13E… (十九; i.e., 19)

Predicted Number (in brackets) was Not Present

2E 3Ch 5Ch 6E 8Ch… (十; i.e., 10)
4E 6E 8Ch 10Ch 12E… (13)
6E 8E 9Ch 11E 12Ch… (16)
9E 11E 12E 14Ch 15Ch… (19)
10E 12Ch 14Ch 16Ch 18Ch… (二十一; i.e., 21)
0E 2E 4Ch 6E 8E… (十三; i.e., 13)
6Ch 9E 12Ch 15Ch 18E… (二十八; i.e., 28)
8Ch 10Ch 11E 13Ch 14E… (21)

Note. Ch = Chinese language stimuli, E = English language stimuli

APPENDIX H: LOW-RISK ETHICS NOTIFICATION FOR STUDY 5



Human Ethics Notification - 4000018234

humanethics@massey.ac.nz <humanethics@massey.ac.nz> Aug 4, 2017 at 2:41 PM
To: A.Lindsay@massey.ac.nz, Martina.Daskova.1@uni.massey.ac.nz, A.P.Gilbey@massey.ac.nz
Cc: M.E.Thomas@massey.ac.nz

HoU Review Group

Ethics Notification Number: 4000018234
Title: Language switching when listening to a conversation in an aircraft cockpit and radio transmissions

Thank you for your notification which you have assessed as Low Risk.
Your project has been recorded in our system which is reported in the Annual Report of the Massey University Human Ethics Committee.

The low risk notification for this project is valid for a maximum of three years.
If situations subsequently occur which cause you to reconsider your ethical analysis, please log on to http://rims.massey.ac.nz and register the changes in order that they be assessed as safe to proceed.

Please note that travel undertaken by students must be approved by the supervisor and the relevant Pro Vice-Chancellor and be in accordance with the Policy and Procedures for Course-Related Student Travel Overseas. In addition, the supervisor must advise the University's Insurance Officer.

A reminder to include the following statement on all public documents:

"This project has been evaluated by peer review and judged to be low risk. Consequently it has not been reviewed by one of the University's Human Ethics Committees. The researcher(s) named in this document are responsible for the ethical conduct of this research.
If you have any concerns about the conduct of this research that you want to raise with someone other than the researcher(s), please contact Dr Brian Finch, Director (Research Ethics), email humanethics@massey.ac.nz. "

Please note that if a sponsoring organisation, funding authority or a journal in which you wish to publish require evidence of committee approval (with an approval number), you will have to complete the application form again answering yes to the publication question to provide more information to go before one of the University's Human Ethics Committees. You should also note that such an approval can only be provided prior to the commencement of the research.

You are reminded that staff researchers and supervisors are fully responsible for ensuring that the information in the low risk notification has met the requirements and guidelines for submission of a low risk notification.

If you wish to print an official copy of this letter, please login to the RIMS system, and under the Reporting section, View Reports you will find a link to run the LR Report.

Yours sincerely
Dr Brian Finch
Chair, Human Ethics Chairs' Committee and
Director (Research Ethics)

APPENDIX I: STIMULI LIST FOR STUDY 5

Task: Call Sign Recognition

L2 Condition	*Mix Condition*
Target	*Target*
867	341
Distracting Stimuli	*Distracting Stimuli*
142	137 Chinese
193	407 Chinese
346	654 Chinese
379	835 Chinese
472	987 Chinese
531	123 English
629	265 English
685	562 English
751	649 English
961	729 English

Task: Error Identification

L2 Condition	*Mix Condition*
Correct	*Correct*
2 + 8 = 10	3 + 2 = 5 Chinese
4 + 5 = 9	4 + 6 = 10 Chinese
5 + 2 = 7	4 – 2 = 2 Chinese
6 + 4 = 10	6 – 1 = 5 Chinese
5 – 2 = 3	2 + 6 = 8 English
4 – 1 = 3	3 + 1 = 4 English
6 – 2 = 4	7 – 2 = 5 English
8 – 7 = 1	9 – 4 = 5 English
Incorrect	*Incorrect*
1 + 4 = 6	3 + 6 = 7 Chinese
7 + 3 = 4	9 – 2 = 1 Chinese
3 – 2 = 0	5 + 3 = 7 English
9 – 5 = 6	8 – 6 = 4 English

Task: Prediction

L2 Condition	*Mix Condition*
Predicted Number (in brackets) was Present	*Predicted Number (in brackets) was Present*
0 3 6 9 12… (15)	4E 6E 8Ch 10Ch 12E… (16)
2 5 8 11 14… (20)	5E 8Ch 11E 14Ch 17Ch… (23)
3 5 7 9 11… (15)	6Ch 9Ch 12E 15E 18E… (21)
8 10 12 14 16… (18)	
Predicted Number (in brackets) was Not Present	*Predicted Number (in brackets) was Not Present*
1 3 5 7 9… (14)	0E 2Ch 4Ch 6E 8Ch… (13)
4 7 10 13 16… (18)	1Ch 4E 7Ch 10E 13Ch… (20)
9 12 15 18 21… (25)	5Ch 7Ch 9E 11E 13E… (16)
11 13 15 17 19… (22)	7Ch 10E 13E 16E 19Ch… (23)
	10Ch 12Ch 14E 16Ch 18E… (21)

APPENDIX J: TRANSCRIPT OF BACKGROUND TALKS FOR STUDY 5

Talk 1 in English

00:12
This is where I live. I live in Kenya, at the south parts of the Nairobi National Park. Those are my dad's cows at the back, and behind the cows, that's the Nairobi National Park. Nairobi National Park is not fenced in the south widely, which means wild animals like zebras migrate out of the park freely. So predators like lions follow them, and this is what they do. They kill our livestock. This is one of the cows which was killed at night, and I just woke up in the morning and I found it dead, and I felt so bad, because it was the only bull we had.

00:59
My community, the Maasai, we believe that we came from heaven with all our animals and all the land for herding them, and that's why we value them so much. So I grew up hating lions so much. The morans are the warriors who protect our community and the livestock, and they're also upset about this problem. So they kill the lions. It's one of the six lions which were killed in Nairobi. And I think this is why the Nairobi National Park lions are few.

01:36
So a boy, from six to nine years old, in my community is responsible for his dad's cows, and that's the same thing which happened to me. So I had to find a way of solving this problem. And the first idea I got was to use fire, because I thought lions were scared of fire. But I came to realize that that didn't really help, because it was even helping the lions to see through the cowshed. So I didn't give up. I continued. And a second idea I got was to use a scarecrow. I was trying to trick the lions [into thinking] that I was standing near the cowshed. But lions are very clever. They will come the first day and they see the scarecrow, and they go back, but the second day, they'll come and they say, this thing is not moving here, it's always here. So he jumps in and kills the animals. So one night, I was walking around the cowshed with a torch, and that day, the lions didn't come. And I discovered that lions are afraid of a moving light. So I had an idea. Since I was a small boy, I used to work in my room for the whole day, and I even took apart my mom's new radio, and that day she almost killed me, but I learned a lot about electronics. So I got an old car battery, an indicator box. It's a small device found in a motorcycle, and it helps motorists when they want to turn right or left. It blinks. And I got a switch where I can switch on the lights, on and off. And that's a small torch from a broken flashlight.

03:34
So I set up everything. As you can see, the solar panel charges the battery, and the battery supplies the power to the small indicator box. I call it a transformer. And the indicator box makes the lights flash. As you can see, the bulbs face outside, because that's where the lions come from. And that's how it looks to lions when they come at night. The lights flash and trick the lions into thinking I was walking around the cowshed, but I was sleeping in my bed.

04:16
So I set it up in my home two years ago, and since then, we have never experienced any problem with lions. And my neighbouring homes heard about this idea. One of them was this grandmother. She had a lot of her animals being killed by lions, and she asked me if I could put the lights for her. And I said, "Yes." So I put the lights. You can see at the back, those are the lion lights. Since now, I've set up seven homes around my community, and they're really working. And my idea is also being used now all over Kenya for scaring other predators like hyenas, leopards, and it's also being used to scare elephants away from people's farms. Because of this invention, I was lucky to get a scholarship in one of the best schools in Kenya, Brookhouse International School, and I'm really excited about this. My new school now is coming in and helping by fundraising and creating an awareness. I even took my friends back to my community, and we're installing the lights to the homes which don't have [any], and I'm teaching them how to put them.

05:37
So one year ago, I was just a boy in the savannah grassland herding my father's cows, and I used to see planes flying over, and I told myself that one day, I'll be there inside. And here I am today. I got a chance to come by plane for my first time for TED. So my big dream is to become an aircraft engineer and pilot when I grow up.

06:04
I used to hate lions, but now because my invention is saving my father's cows and the lions, we are able to stay with the lions without any conflict. Ashê olên. It means in my language, thank you very much.

Talk 1 in Chinese

00:12
这是我所生活的地方，肯尼亚， 在内罗毕国家公园的南部。 在我的身后是我父亲的奶牛， 在奶牛的后边 就是内罗毕国家公园。 内罗毕国家公园的南部并没有全部围起栅栏， 这就意味着像斑马这样的野生动物 可以自由地在公园外移动， 所以像狮子这样的捕食者，会跟随他们。 这就是他们的所作所为， 他们捕杀了我们的家畜。 这就是其中一只在夜晚被杀害的奶牛， 我在早晨醒来时就发现它已经死了， 我的心情槽透了， 因为这是我家里唯一的公牛。

00:59
我们马赛族的人相信， 我们是带着我们的动物和家园从天堂而来， 然后放牧、生活，这就是我们为什么如此重视它们， 所以我逐渐地变得十分厌恶狮子。 莫兰人都是勇士， 他们保护我们的家族和牲畜， 但是他们同样对这个问题感到沮丧，素手无策 所以他们决定杀了这些狮子。 这是他们在内罗毕国家公园杀的六个狮子中的其中一个， 我想这就是内罗毕公园的狮子会那么少的原因。

01:36
在我们家族，凡是六到九岁的男孩 都有肩负着保卫他们父亲奶牛的责任， 同样的事业发生在我身上。所以我必须找到一个解决问题的方法。 我想到的第一个办法是使用火， 因为我知道狮子们都惧怕火。 但我又意识到这并没有什么成效，反倒还帮助了狮子 看到了我们的牛棚。 但我并没有因此放弃，我坚持不懈

着 于是有了第二个办法： 利用稻草人。 我想要欺骗狮子 让它们误以为是我站在牛舍旁边。 但是狮子十分聪明。 他们第一天来时看到了稻草人，然后就回去了。 但是第二天，他们会再来并且说： 这个东西从没动过，他一直呆在这儿 所以他们又跳进来，杀死了动物。 有一天晚上，我拿着手电筒在牛棚边走动， 那天晚上，狮子并没有来。 我开始意识到狮子们会害怕移动的光。 所以我又有了个主意。 因为我当时是个小孩子， 我可以一整天不干别的，在自己的房间里鼓捣一天， 我甚至把我妈妈新买的收音机拆得七零八落， 那天她差点杀了我， 尽管如此，我学到了许多关于电子的知识。 所以我找到了一个旧的车用蓄电池， 一个指示器。这是一个从摩托车中找到的小装备， 用来控制摩托车的转向灯，能够让灯闪烁。 同时我找到了一个开关，可以控制灯的亮灭。 灯泡来自于一个坏掉的手电筒。

03:34
一切都准备好了。 就如你所见，由太阳能接收板来给电池充电， 再由电池来指示器给提供能量 我把这个叫做“变压器”， 这个指示器又会使光线闪动， 你可以看到，这个电灯泡是朝外的， 因为狮子是从那儿过来的。 当狮子晚上走近的时候看到的就是这个样子。 这个灯光一闪动就会使狮子受到欺骗 它们会以为是我在牛棚附近走动， 但实际上，此时的我正在自己的床上做着春秋大梦。

04:16
我在两年前把这个装置装在了我家中，从那时起，我们就再没有狮子捕食家禽的烦恼了。 我的邻居们都听说了这个办法。 他们中的其中一个就是这位老奶奶。 她家有一大堆的家禽被狮子捕食了。 她问我是否可以帮她在家里也装一个这样的灯。 我欣然同意。 我装上了这些灯，你们可以在背后看到，这些就是赶狮灯。 至今为止，我已为我们家族社区中的七户家庭， 装上了这个赶狮灯， 它们非常有用。 现在，我的方法已经被肯尼亚的所有人使用 来驱赶其他类似的捕食动物，像鬣狗，美洲豹等。 它们同样被用来 驱赶大象远离人类的农庄。我也很幸运地因为这项发明而获得了奖学金， 这是肯尼亚最好的学校—— 布鲁克豪斯国际学校所颁发的， 我对此感到十分的兴奋。 我的新学校现在也加入进来，并帮助我 提高知名度并筹集更多资金。 我甚至将我的朋友带回我的社区， 帮助他们在没有赶狮灯的人家里安装这个。 我教他们怎么安装。

05:37
一年前，我只是一个热带大草原上， 帮爸爸放牛的普通男孩， 我常常看着飞机飞过， 并且告诉自己：总有一天，我也会在里面。 今天，我做到了。 我获得了乘飞机来 TED 的机会，这是我的第一次 所以我伟大的梦想是：当我长大以后，成为一个飞机工程师兼飞行员。

06:04
我过去讨厌狮子，但是现在，就是因为我的发明 拯救了我爸爸的牛， 以及狮子， 我们终于可以狮子和平共处了。 Ashê olên。在我的语言里，它的意思是：非常感谢！

Talk 2 in English

00:11
My name is Joseph, a Member of Parliament in Kenya. Picture a Maasai village, and one evening, government soldiers come, surround the village and ask each elder to bring one boy to school. That's how I went to school -- pretty much a government guy pointing a gun and told my father, "You have to make a choice." I walked very comfortably to this missionary school that was run by an American missionary. The first thing the American missionary gave me was a candy. I had never in my life ever tasted candy. So I said to myself, with all these hundred other boys, this is where I belong.

00:42
I stayed. When everybody else was dropping out. My family moved; we're nomads. It was a boarding school, I was seven -- Every time it closed you had to travel to find them. 40-50 miles, it doesn't matter. You slept in the bush, but you kept going.

00:57
And I stayed. I don't know why, but I did. All of a sudden I passed the national examination, found myself in a very beautiful high school in Kenya. And I finished high school. And just walking, I found a man who gave me a full scholarship to the United States. My mother still lived in a cow-dung hut, none of my brothers were going to school, and this man told me, "Here, go."

01:19
I got a scholarship to St. Lawrence University, Upstate New York; finished that. And after that I went to Harvard Graduate School; finished that. Then I worked in DC a little bit: I wrote a book for National Geographic and taught U.S. history. And every time, I kept going back home, listening to their problems -- sick people, people with no water, all this stuff -- every time I go back to America, I kept thinking about them.

01:45
Then one day, an elder gave me a story that went like this: long time ago, there was a big war between tribes. This specific tribe was really afraid of this other Luhya tribe. Every time, they sent scouts to make sure no one attacked them. So one day, the scouts came running and told the villagers, "The enemies are coming. Only half an hour away, they'll be here." So people scrambled, took their things and ready to go, move out. But there were two men: one man was blind, one man had no legs -- he was born like that. The leader of the chiefs said, "No, sorry. We can't take you. You'll slow us down. We have to flee our women and children, we have to run." And they were left behind, waiting to die.

02:26
But these two people worked something out. The blind man said, "Look, I'm a very strong man but I can't see." The man with no legs says, "I can see as far as the end of the world, but I can't save myself from a cat, or whatever animals." The blind man went down on his knees like this, and told the man with no legs to go over his back, and stood up. The man on top can see, the blind man can walk. These guys took off, followed the footsteps of the villagers until they found and passed them.

02:58
So, this was told to me in a setup of elders. And it's a really poor area. I represent Northern Kenya: the most nomadic, remote areas you can even find. And that man told me, "So, here you are. You've got a good education from America, you have a good life in America; what are you going to do for us? We want you to be our eyes, we'll give you the legs. We'll walk you, you lead us."

03:24
The opportunity came. I was always thinking about that: "What can I do to help my people? Every time you go to an area where for 43 years of independence, we still don't have basic health facilities. A man has to be transported in a wheelbarrow 30 km for a hospital. No clean drinking water.

03:39
So I said, "I'm going to dedicate myself. I'm leaving America. I'm going to run for office." Last June, I moved from America, ran in July election and won. And I came for them, and that's my goal.

03:57
Right now I have in place, for the last nine months, a plan that in five years, every nomad will have clean drinking water. We're building dispensaries across that constituency. I'm asking my friends from America to help with bringing nurses or doctors to help us out. I'm trying to improve infrastructure. I'm using the knowledge I received from the United States and from my community to move them forward. I'm trying to develop home grown solutions to our issues because people from outside can come and help us, but if we don't help ourselves, there's nothing to do.

04:33
My plan right now as I continue with introducing students to different fields -- some become doctors, some lawyers -- we want to produce a comprehensive group of people, students who can come back and help us see a community grow that is in the middle of a huge economic recession.

04:49
As I continue to be a Member of Parliament and as I continue listening to all of you talking about botany, health, democracy, new inventions, I'm hoping that one day in my own little community -- which is 26,000 square km, maybe five times Rhode Island -- with no roads, we'll be able to become a model to help others develop. Thank you very much.

Talk 2 in Chinese

00:11
我叫约瑟夫，是肯尼亚的一名议员。 想像一下，有一天晚上， 在一个马塞村庄里，来了一队政府军，他们围起村子， 要每个家长都送一个男孩去上学。 我是这样开始上学的—— 一个兵用枪指着我爸 说：“你必须得做个决定。” 那是个教会学校，一个美国传教士开的， 当我相当惬意地走进学校时， 这位传教士递给我的第一件东西居然是一块糖 我以前从来都没有尝过糖是什么滋味儿 在几百名

男孩面前，我心想到， 这儿才是我属于的地方。 别人退学了，我一直留下来。 我的家庭迁来迁去。我们是游牧人。 每次学期结束学校放假——那是个寄宿学校，我那时七岁—— 我就得一路穿行直到找到他们为止。 50 英里，40 英里，都不重要。 睡在灌木丛里，可是你还是得继续走。

00:56
而我坚持到底。不知道为什么我会坚持，但是我就是待了下去。 然后突然间，我通过了全国中考， 发现自己进入了一个肯尼亚的美丽的高中。 我读完了高中。然后我就这么走着，发现有一个人 给了我一笔去美国的全额奖学金。 我妈妈仍旧住在遍地牛粪的棚屋里， 我的兄弟们没有一个在上学， 而这个人却对我说，“给你，去吧。”

01:19
于是我拿着奖学金去了纽约州北部的圣劳伦斯大学。 念完本科后，我又去了哈佛读研。 读完研后，我在华盛顿工作了一阵子。 我给国家地理写了一本书，也教授美国历史的课程。 我每次回到我的家乡， 不断的听到人们的问题， 人们患病，人们没有水，所有这种事。 而且我每次回到美国，都不断的思考这些问题。

01:44
后来有一天，一位长辈给我讲了一个故事，是这样说的—— 很久以前，部落之间发生了一场大战。 其中有一个部落，特别地害怕另一个卢赫雅部落。 他们每次都会派人去侦察，以确保没有人来袭击自己。 有一天，侦查员跑回来告诉村民们， 说：“敌人们要来了，还只有半个小时，他们就会到这儿。” 于是人们乱作一团，拿好自己的东西准备撤， 但是其中却有这样两个男人， 一个是眼盲了，一个天生没有双腿。 酋长说到：“不，对不起，我们不能带你们走，你们会拖大家的后腿。 我们得让我们的妇女和儿童们逃走，我们必须得跑。” 于是他俩被留下等死。

02:26
但是这两个人却想了一个办法。 眼盲的男人说：“喏，我是个很强壮的人，但是我看不见东西。” 没有双腿的人说道：“我能看见远至世界的尽处， 但是我却敌不过任何动物，乃至一只猫。” 于是盲人蹲了下来，就像这样， 他告诉无腿的人骑到自己背上，然后站了起来。 骑在上面的人能看，而盲人能走。 这俩人便沿着村民们留下的脚印一路出发了， 直到他们赶上了其他人，并且还超越了他们。

02:58
我听着故事的时候，周围还有其他长辈们。 那是个相当贫困的地方，我指的是肯尼亚北部—— 你能找得到的，最偏远最流浪的地方。 讲故事的人对我说：“你在美国受过良好的教育， 也在那儿过着优越的生活， 现在你来我们这儿，你能为我们做些什么呢？ 我们希望你当我们的眼睛，我们就做你的腿。 你领着我们大家，我们举着你走。”

03:23
于是机遇就这么来了，我总是在思索， 为了帮助我的人民，我能做些什么？ 每当你来到这个43年前就已经独立的国家， 我们仍然还是没有基本卫生设施。 一个病人得用手推车送去20、30公里外的医院就诊，没有干净的饮用水。

03:38
于是我说，“我要献一份力， 我要离开美国。我要回去参加竞选。” 因此去年，我六月里离开的美国，在七月参加竞选并当选。 我是为他们才来，这是我的目标。

03:56
到如今，我在这位子上已有九个月了， （我）计划五年内，让每个游牧人都能用喝上干净的饮用水。 我们会在选区里建立诊所。 我向我的美国朋友们寻求帮忙 带来护士和医生们帮助大家摆脱困难。 我尝试改善基础设施。 我应用从美国以及从我自己社区中学到的知识 来促进实施。 对于大家面临的问题，我试图制定出自己的解决方案。 因为我们意识到外面的人能过来帮助我们， 但是如果我们连自己都不帮自己的话，那我们真就没什么指望了。

04:33
所以我现在的计划就是，随着我继续引进各个领域的学生们—— 其中有一些会成为医生，一些会是律师—— 我们想建立一个全面综合的群体， 这些学生们会回来帮助大家 见证经济大衰退中一个社会的成长。

04:49
因此，当我还是议会议员， 当我不断的听你们谈论着植物学， 谈论健康，谈论民主，谈论新发明， 我就希望能有一天，在我自己的社区里—— 它占地两万六千平方公里， 大约相当于五个罗德岛的面积，还没有公路—— 我们能成为帮助别人发展的模范。 非常感谢!

Talk 3 in English

00:11

In 1827, a fellow called George Pocock actually pioneered the use of kites for towing buggies in races against horse carriages across the English countryside. Then of course, at the dawn of aviation, all of the great inventors of the time -- like Hargreaves, like Langley, even Alexander Graham Bell, inventor of the telephone, who was flying this kite -- were doing so in the pursuit of aviation.

01:19
Then these two fellows came along, and they were flying kites to develop the control systems that would ultimately enable powered human flight. So this is of course Orville and Wilbur Wright, and the Wright Flyer. And their experiments with kites led to this momentous occasion, where we powered up and took off for the first-ever 12-second human flight. And that was fantastic for the future of commercial aviation.

01:45
But unfortunately, it relegated kites once again to be considered children's toys. That was until the 1970s, where we had the last energy crisis. And a fabulous man called Miles Loyd who lives on the outskirts of San Francisco, wrote this seminal paper that was completely ignored in the Journal of Energy about how to use basically an airplane on a piece of string to generate enormous amounts of electricity. The real key observation he made is that a free-flying wing can sweep through more sky and generate more power in a unit of time than a fixed-wing turbine.

02:18
So turbines grew. And they can now span up to three hundred feet at the hub height, but they can't really go a lot higher, and more height is where the more wind is, and more power -- as much as twice as much.

02:29
So cut to now. We still have an energy crisis, and now we have a climate crisis as well. You know, so humans generate about 12 trillion watts, or 12 terawatts, from fossil fuels. And Al Gore has spoken to why we need to hit one of these targets, and in reality what that means is in the next 30 to 40 years, we have to make 10 trillion watts or more of new clean energy somehow. Wind is the second-largest renewable resource after solar: 3600 terawatts, more than enough to supply humanity 200 times over. The majority of it is in the higher altitudes, above 300 feet, where we don't have a technology as yet to get there.

03:10
So this is the dawn of the new age of kites. This is our test site on Maui, flying across the sky. I'm now going to show you the first autonomous generation of power by every child's favourite plaything. As you can tell, you need to be a robot to fly this thing for thousands of hours. It makes you a little nauseous. And here we're actually generating about 10 kilowatts -- so, enough to power probably five United States households -- with a kite not much larger than this piano. And the real significant thing here is we're developing the control systems, as did the Wright brothers that would enable sustained, long-duration flight. And it doesn't hurt to do it in a location like this either.

Talk 3 in Chinese

00:11
在1827年，一个叫做乔治朴考克的家伙 实际上先锋般的用风筝来牵动车辆 这样来和马匹做横跨英国的比赛。 随后，理所当然的， 在航空时代即将到来之际， 那个时代所有的伟大发明家们-- 像哈格里夫斯，像兰利， 甚至是亚历山大格雷厄姆贝尔，电话的发明着，都会像这样放风筝-- 这样做是为了追求航空能力。

01:19
然后这两个家伙出现了， 他们在风筝上开发了制动系统 那样会最终让人类的飞行梦想成真。 这当然是恩奎斯特和莱特兄弟， 以及莱特飞行器。 他们对风筝的实验引发了这个 重要时刻，那就是我们可以从起飞到降落 第一个12秒的人类飞行。 而且那是一个美妙的商业航空的未来。

01:45
但是不幸的是，风筝再一次被低估为儿童的玩具。 直到 19 世纪 70 年代，当我们上一次经济危机的时候。 一个叫做迈尔斯劳埃德的神话般的人 他住在旧金山的郊区。 写了这样一篇完全被遗忘的论文， 能量之旅， 这是关于怎样用一根绳子去操控飞机 来产生超乎强大的电力。 他做到的最关键的观察是 一个自由飞行翼可以扫过更多的天空和创造更多的能量 这是在一定的时间内和固定翼涡轮相比。

02:18
所以涡轮那时在不断的被开发。现在他们可以跨越跨度长达三百英尺的枢纽高度， 但是他们确实在也不能比那再高了， 不过更高的地方有更多的风，以及更多的能量 -- 两倍高的能量。

02:29
所以直到现在，我们仍然还有能量危机， 现在我们还同时有环境危机，你知道的。 所以人类生产了大概十二万亿瓦特， 或 12 兆瓦，从化石燃料中。 然后戈尔已经讲解了我们为什么要达到这些目标， 以及在今后的 30 到 40 年里那到底意味着什么， 我们必须制造 10 亿瓦特或者更多的干净能源，不管怎么样都要。 风能是在太阳能之后的第二大可再生能源： 3600 兆瓦特，足以维持比现在多 200 倍的人。 大多数风能是在高海拔，300 英尺以上， 我们现有的科技还无法到达那个高度。

03:10
这样就到了风筝新舞台开始的时候。 这是我们在茂宜岛上的一个测试点，在空中翱翔着。 我现在将要为你们展示 第一自主发电机 由所有小孩最爱的玩具所提供。 你可以想象，你需要一个机器人来几千小时的放飞这个东西。 这可能让你有点眩晕。 而且这样我们实际上可以生产大概 10 千瓦-- 所以，足够维持 5 个美国家庭用电-- 就用一个不大于钢琴大小的风筝。 不过最有意义的事情是 我们正在开发控制系统， 就像莱特兄弟那样，这样会提供可持续，长时间的飞行。 而也不会去破坏像这样的环境。