UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING AND THE ENVIRONMENT

Institute of Sound and Vibration Research

Developing a measure of auditory fitness for duty for military personnel

by

Hannah Domenica Semeraro

Thesis for the degree of Doctor of Philosophy

Submission of soft-bound thesis: December 2015

UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING AND THE ENVIRONMENT

Audiology

Thesis for the degree of Doctor of Philosophy

DEVELOPING A MEASURE OF AUDITORY FITNESS FOR DUTY FOR MILITARY PERSONNEL

Hannah Domenica Semeraro

ABSTRACT

The ability to listen to and understand commands in noisy environments, whilst maintaining situational awareness, is an important skill for military personnel, and can be critical for mission success. Due to the nature of their work, military personnel are regularly exposed to damaging noise exposures, which could lead to hearing loss and therefore the inability to understand commands. Accurately measuring auditory fitness for duty (AFFD) is important for ensuring that personnel have sufficient hearing ability to be effective in operational scenarios. Pure-tone audiometry (PTA) is the hearing test currently used by the UK military but it is not known whether it is able to accurately predict AFFD. The aims of this thesis were to: 1) better understand AFFD, focusing on the infantry; and 2) undertake the initial development of a credible alternative to PTA and a simulation of an AFFD task, ahead of future research to determine which test(s) best predict AFFD.

Using focus groups, followed by a questionnaire, 17 mission-critical auditory tasks (MCATs) carried out by infantry personnel were identified. Nine of these tasks were prioritised for evaluating AFFD and seven were speech communication tasks (SC-MCATs). It was anticipated that a speech-in-noise test might be a better tool than PTA for predicting performance on the SC-MCATs. Following a review of existing speech-in-noise tests, including a consultation with military subject-matter experts, the Coordinate Response Measure (CRM) was selected for this purpose, partly due to the high face validity when compared to typical infantry command structure. The CRM speech material was re-recorded in British English using NATO call-signs, was equalised in terms of intelligibility, and was implemented into an adaptive procedure with stationary speech-spectrum noise and evaluated using normal hearing civilians and hearing impaired military personnel. An AFFD task simulation of the SC-MCATs was developed. It simulated the environment of listening to commands over a military radio in a moving armoured vehicle. A final study found that while both the AFFD task simulation and the CRM were adversely affected by simulated hearing loss, only the task simulation appeared to be affected by experience of military commands. Further work is now required to determine whether PTA or the CRM, when combined with additional information such as previous military experience, best predict performance on the AFFD task simulation.

Table of Contents

Table of Contents ..... **i**

List of Tables ..... **vii**

List of Figures ..... **xi**

DECLARATION OF AUTHORSHIP ..... **xv**

Acknowledgements ..... **xvii**

Glossary **xix**

Abbreviations ..... **xx**

Chapter 1: Introduction **1**

1.1 Background and motivation 1

1.2 Thesis structure, research objectives and dissemination of findings 3

1.3 Contributions to knowledge 6

Chapter 2: Auditory fitness for duty (AFFD) in the military **9**

2.1 Introduction 9

2.2 Situational awareness 9

2.3 Hearing loss in the military 11

2.3.1 Introduction 11

2.3.2 Occupational noise exposure 12

2.3.3 Military noise exposure 13

2.3.4 Noise induced hearing loss 14

2.3.5 Additional causes of military hearing loss 17

2.3.6 Summary 18

2.4 Auditory fitness for duty 18

2.4.1 Introduction 18

2.4.2 Hearing critical tasks 18

2.4.3 Auditory fitness for duty assessment and challenges 20

2.4.4 Summary 23

2.5 Military auditory fitness for duty assessment 23

2.5.1 Introduction 23

2.5.2 Armed Forces hearing assessment protocol 24
2.5.3 Summary 27
2.6 Chapter 2 summary 27
Chapter 3: Identification of mission-critical auditory tasks (MCATs) 29
3.1 Introduction 29
3.2 Research objective 1 29
3.3 Study 1 part A: exploring auditory tasks 30
3.4 Study 1 part B: identification of MCATs 35
3.4.1 Introduction 35
3.4.2 Aims 36
3.4.3 Methods 36
3.4.4 Results: consequences of poor performance 40
3.4.5 Results: roles that carry out each task 41
3.4.6 Results: frequency of task performance 42
3.4.7 Results: identifying tasks to be represented in a measure of AFFD 42
3.4.8 Results: test-retest reliability 45
3.4.9 Discussion 45
3.4.10 Conclusion 48
3.5 Chapter 3 summary 49
Chapter 4: Is pure-tone audiometry (PTA) fit for duty? 51
4.1 Introduction 51
4.2 PTA: A measure of hearing acuity not hearing ability 52
4.2.1 Psychoacoustic abilities contributing to speech intelligibility 54
4.2.2 Non-psychoacoustic factors influencing speech intelligibility 60
4.2.3 Summary 61
4.3 An approach for exploring the predictive validity of PTA as a measure of performance on the speech communication MCATs 62
4.4 Reporting the association between PTA and speech intelligibility tests 65
4.5 Chapter 4 Summary 70

Chapter 5: Developing a new auditory fitness for duty measure..... 71

5.1 Introduction 71

5.2 Speech-in-noise testing: reviewing and selecting an appropriate test 72

5.2.1 Introduction 72

5.2.2 Overview of available speech-in-noise tests 72

5.2.3 Considerations when selecting a speech-in-noise test to measure military AFFD..... 74

5.2.4 Pilot study: rating the similarity of speech test stimuli to command structure 81

5.2.5 Chosen speech-in-noise test: the coordinate response measure (CRM) 82

5.3 Study 2: Developing and recording CRM speech-in-noise test 83

5.3.1 Introduction 83

5.3.2 Justification for re-recording the CRM 83

5.3.3 Research objective 2..... 84

5.3.4 Pilot study 1: deciding the format for recording sentences..... 84

5.3.5 Recording the CRM stimuli 85

5.3.6 Pilot study 2: evaluation and selection of CRM stimuli..... 86

5.3.7 Summary..... 87

5.4 Study 3: Equalising the intelligibility of the CRM in noise 88

5.4.1 Introduction 88

5.4.2 Research objective 3..... 90

5.4.3 Method (route mean square equalised: sessions one and two)..... 90

5.4.4 Results (sessions one and two)..... 92

5.4.5 Method (amplitude equalised: session three) 98

5.4.6 Results (session three)..... 98

5.4.7 Discussion 104

5.4.8 Conclusion..... 106

5.5 Study 4: Exploring the measurement precision of the CRM implemented in an adaptive procedure..... 107

5.5.1 Introduction 107

5.5.2 Research objective 4 and Study 4 aims 113
5.5.3 CRM adaptive procedure characteristics 113
5.5.4 Methods 117
5.5.5 Results: overview 122
5.5.6 Results: stability 124
5.5.7 Results: variability 129
5.5.8 Results: concurrent validity 131
5.5.9 Results: test condition comparisons 135
5.5.10 Discussion and conclusions 137
5.6 Chapter 5 Summary 143

Chapter 6: Developing a method to assess performance on the speech communication MCATs 145

6.1 Introduction 145
6.2 Methods for measuring performance on the SC-MCATs 147
6.2.1 Live training scenarios 148
6.2.2 Simulated training scenarios 149
6.2.3 Task simulation in a clinical environment 151
6.2.4 Chosen method: justification and limitations 152
6.2.5 Summary 154
6.3 Study 5: Developing, recording and evaluating the Vehicle Communication Simulated MCAT (VEHCOM-SimMCAT) test 155
6.3.1 Introduction 155
6.3.2 Research objective 5 155
6.3.3 VEHCOM SimMCAT development 156
6.3.4 Finalised VEHCOM SimMCAT 165
6.4 Study 6: Initial assessment of the VEHCOM SimMCAT: is it sensitive to simulated hearing impairment and military experience? 167
6.4.1 Introduction 167
6.4.2 Research objective 6 and Study 6 aims 170
6.4.3 Hearing loss simulation 170
6.4.4 Method 180

6.4.5 Results 183
6.4.6 Discussion and conclusions 189
6.4.7 Moving towards assessing the predictive validity of the CRM as a measure of AFFD: next steps 195
6.5 Chapter 6 Summary 197

Chapter 7: Summary, conclusions and future research 199

7.1 Summary 199
7.2 Conclusions 201
7.3 Future work 202

Appendices 205
Appendix A Types of validity 207
Appendix B Background to psychometric functions (PF) 209
Appendix C Survey to obtain the opinions of military personnel on speech tests 213
Appendix D MATLAB code for generating stationary speech-spectrum noise . 215
Appendix E MATLAB code equalising the RMS of the CRM speech stimuli 217
Appendix F MATLAB code CRM method of constant stimuli 219
Appendix G Graphs showing direct comparison of psychometric functions before and after intelligibility equalisation 221
Appendix H Graphs showing Bland and Altman plots for between session and repeat changes in SRT score 223
Appendix I Visual awareness tasks development- 'Tag the Enemy' 229
Appendix J List of recorded commands for each speech communication MCATs 233
Appendix K Pilot Experiment: selecting an SNR for the VEHCOM SimMCAT ... 235
Appendix L List of MATLAB Code Authors 237
Appendix M Confirmation of ethics approval 239
Appendix N List of presentations 245
Appendix O Participant recruitment challenges and strategies in the military population 247
References 249

List of Tables

Table 2.1 European Commission (2008) guidelines on noise exposure regulations 12

Table 2.2 Armed Forces hearing acuity grades (MoD, 2013) 26

Table 2.3 Age and gender specific hearing threshold assessment (MoD, 2013).....26

Table 3.1 List of focus group questions 34

Table 3.2 Focus groups: themes and sub-themes 34

Table 3.3 Focus group: theme one- infantry auditory tasks.....34

Table 3.4 Questionnaire survey guide 39

Table 3.5 Study 1 part B participant information 39

Table 3.6 Percentage of responses for each task for each consequence rating. The shaded black section shows the median rating. Each column totals 100%. 40

Table 3.7 Percentage of responses for each task for the question 'who performs this task'. The shaded black section shows the median rating. Each column totals 100%. ...41

Table 3.8 Infantry personnel reported to perform tasks 2, 4 and 8; the three tasks with the highest number of responses for 'some infantry personnel'. Number in brackets indicates the number of responses given for each role 41

Table 3.9 Percentage of responses for each frequency rating. The shaded black section shows the median rating. Each column totals 100%. 42

Table 3.10 List of MCATs to be prioritised for representation in a measure of AFFD for infantry personnel 45

Table 4.1 Literature reporting correlations between PTA and speech intelligibility tests in quiet (SQ) or noise (SN) with normal hearing (NH) and hearing impaired (HI) listeners 68

Table 4.2 Overview of methods used in papers listed in Table 4.1 69

Table 5.1 Overview of a selection of available speech tests 73

Table 5.2 Quotes from Study 1 Part A. Infantry and combat support personnel's opinions on pure-tone audiometry 80

Table 5.3 Results of survey to investigate face validity of speech-in-noise tests for military AFFD testing 81

Table 5.4 List of stimuli in the University of Southampton recording of the British English Coordinate Response Measure (CRM) 84

Table 5.5 Thresholds (SRT 50) for the CRM target word (highlighted red= amplitude altered before session three).....97

Table 5.6 Threshold (SRT 50) scores, threshold 95% confidence intervals and slope for each CRM target word 101

Table 5.7 Target word group mean SRT 50 score details for sessions 1 & 2 (combined) and session 3 102

Table 5.8 Gender and age sample characteristics of participants in Study 4 118

Table 5.9 Study 4 test conditions for normal hearing (NH) and hearing impaired (HI) participants 120

Table 5.10 Study 4 CRM test condition abbreviations 122

Table 5.11 Shapiro-Wilk test of normality across all the data 124

Table 5.12 Normal hearing stability values for CRM adaptive procedure test conditions and TDT, showing the mean change in SRT between repeats and the 95% confidence interval of the mean. 126

Table 5.13 Normal hearing three-way (condition, session and repeat) repeated measures ANOVA results 127

Table 5.14 Hearing impaired stability values for CRM adaptive procedure test conditions and TDT, showing the mean change in SRT between repeats and the 95% confidence interval of the mean. 128

Table 5.15 Hearing impaired two-way (condition and session) repeated measures ANOVA results 129

Table 5.16 Variability (shown as the 95% confidence interval of the true SRT score for any one measurement value) for the CRM adaptive procedure conditions and the TDT 130

Table 5.17 Concurrent validity between the CRM test conditions and the TDT. Pearson's parametric correlation test (r) has been calculated for the hearing impaired data. Spearman's rho (r_s) has been calculated for the normal hearing and combined normal hearing and hearing impaired data. 132

Table 5.18 Concurrent validity between the CRM test conditions (SRT averaged across repeats) and PTA (average pure-tone thresholds for better hearing ear) for the hearing impaired data, (Pearson's r). 133

Table 5.19 Comparison of correlations of CRM test conditions with the TDT and PTA..... 134

Table 5.20 Comparing the SRT scores for normal hearing listeners on the TDT from Study 4 with scores from previous literature 134

Table 5.21 Comparing the SRT scores for normal hearing and hearing impaired listeners on the CRM (CSoff and CSon) and TDT 135

Table 5.22 Measurement precision summary for each of the CRM test conditions 136

Table 5.23 Replicability of the CRM for the hearing impaired sample, both within (shaded black) and between (not shaded) test conditions 137

Table 6.1 Speech communication MCATs 145

Table 6.2 The final command lists and key words/phrases (bold) to be used in the VEHCOM SimMCAT 166

Table 6.3 Description of how the hearing loss simulator (HLS) works 173

Table 6.4 Details of the how the hearing thresholds chosen for simulation fall within the H grades 177

Table 6.5 Comparison of mean SRTs on the CRM (call sign off) for hearing impaired individuals and normal hearing individuals listening through the hearing loss simulator 179

Table 6.6 Description of Study 6 method stages 181

Table 6.7 Overview of averaged results from Study 6 for both the CRM and VEHCOM SimMCAT for both the military and civilian samples 184

Table 6.8 Shapiro-Wilk test of normality for Study 6 data 185

Table 6.9 Comparing CRM-CSoff results between hearing impaired listeners and normal hearing individuals listening through the hearing loss simulator 187

Table 6.10 Independent sample t-tests showing the difference between the VEHCOM SimMCAT scores of the military and civilian populations for each hearing acuity group189

List of Figures

Figure 2.1 The three levels of situational awareness outlined by Endsley (1995) 10

Figure 2.2 Flow diagram showing generic AFFD testing methods.....21

Figure 3.1 Consequence/frequency matrix (Key *speech communication, ~sound localisation, $sound detection), numbers relate to tasks in Table 3.3 43

Figure 3.2 Percentage of responses for no/minor consequence if the task is performed poorly and moderate/major/critical consequence if the task is performed poorly. Error bars show 95% confidence intervals. 44

Figure 3.3 Percentage of responses for tasks performed seldom or yearly/occasionally or monthly and those performed regularly or weekly/frequently or daily/continuously or several times per day. Error bars show 95% confidence intervals. 44

Figure 4.1 Schematic representations of a normal hearing (left) and a hearing impaired (right) auditory filter 55

Figure 4.2 Schematic representation of the temporal information contained in the long-term properties of the temporal envelope and short-term fluctuations of a speech signal 57

Figure 4.3 Diagram showing the concept of loudness recruitment proposed by Moore and Glasberg (1993).....59

Figure 4.4 Diagram explaining how the literature reporting the correlations between PTA and speech intelligibility testing can be used to predict the suitability of PTA as a measure of AFFD.....63

Figure 5.1 Equipment set up for recording CRM sentences 86

Figure 5.2 A screen shot of the CRM graphical user interface 91

Figure 5.3 The CRM sentence compilation 92

Figure 5.4 Examples of a poor logistic function fit (left, data for target word 'six') and good logistic function fit (right, data for target word 'Victor'), averaged across participants from session one.....93

Figure 5.5 Logistic functions for call sign target words (obtained from the mean scores at each SNR from sessions one, n=20, and two, n=18) 94

Figure 5.6 Logistic functions for colour target words (obtained from the mean scores at each SNR from sessions one, n=20, and two, n=18).....94

Figure 5.7 Logistic functions for number target words (obtained from the mean scores at each SNR from sessions one, n=20, and two, n=18) 95

Figure 5.8 Logistic functions for call sign target words (obtained from the mean scores at each SNR from session three) 99

Figure 5.9 Logistic functions for colour target words (obtained from the mean scores at each SNR from session three).....99

Figure 5.10 Logistic functions for number target words (obtained from the mean scores at each SNR from session three) 100

Figure 5.11 Graph showing the relationship between the 95% confidence interval of threshold estimation and the slope value of each target word 101

Figure 5.12 Confusion matrices for the CRM target words across all three sessions (top: call signs, letters refer to first letter of call signs, bottom left: colours, bottom right: numbers). Each column totals 100%, calculated as a percentage of all incorrect responses across all presentations across all Study 3 sessions..... 103

Figure 5.13 Mind map showing the relationship between different aspects of measurement precision 108

Figure 5.14 Example adaptive procedure response plot with labelled key features. 116

Figure 5.15 Hearing acuity distribution of participants in Study 4..... 118

Figure 5.16 Box plots showing the distribution of the SRTs for different levels of hearing acuity for each CRM conditions and the TDT (outliers exceed 1.5 times the interquartile range) 123

Figure 5.17 Normal hearing sample mean SRTs for each repeat across the three test conditions. Error bars display 95% confidence interval of the mean. 125

Figure 5.18 Hearing impaired sample mean SRTs for each repeat across four test conditions. Error bars display 95% confidence interval of the mean. 128

Figure 5.19 Correlations between CRM-CSoff and TDT for normal hearing and hearing impaired data. Normal hearing =◯ and hearing impaired = □ 131

Figure 5.20 Correlations between CRM-CSon and TDT for normal hearing and hearing impaired data. Normal hearing =◯ and hearing impaired = □ 131

Figure 6.1 Diagram showing the relationship between a gold standard measure and an achievable measure of performance on the speech communication MCATs (SC-MCATs) and how a clinical speech-in-noise test may be used to predict performance on either of these. Numbers relate to description in text directly below figure.146

Figure 6.2 Dismounted Close Combat Trainer (DCCT) (Espaillat and Smith, 2010) 150

Figure 6.3 Combined Arms Tactical Trainer (CATT), inside a tank simulator (Wallace, 2012) .. 150

Figure 6.4 The compromise to be made when selecting a measure of military communication performance between a test with a high ecological validity and a test which is simple to develop, run and control. 152

Figure 6.5 Flowchart showing the development stages for the VEHCOM SimMCAT 156

Figure 6.6 (right) Recording set-up. Speaker in anechoic chamber 159

Figure 6.7 (below) Recording set-up. Diagram of room layout and equipment 159

Figure 6.8 (bottom left) Recording set-up. KEMAR set-up in listening room 159

Figure 6.9 A schematic diagram of the positions within Warrior armoured vehicle, showing the commander position where the microphone was placed to record the engine noise.....160

Figure 6.10 Flow-diagram outlining the stages of the MATLAB code used to prepare the command lists ready for presentation to participants. 163

Figure 6.11 A schematic diagram representing the audio file presented to a participant listening to one of the command lists 164

Figure 6.12 Explanation of how the hearing loss simulation applies the effects of amplitude reduction and loudness recruitment to an audio file.....174

Figure 6.13 Average air conduction hearing thresholds of 400 patients receiving care at the Defence Audiology Service 176

Figure 6.14 Hearing threshold chosen to be simulated for Study 6 176

Figure 6.15 Comparing simulated hearing impairments and those seen in the military population 177

Figure 6.16 Comparing performance on the CRM for hearing impaired individuals and normal hearing individuals listening through the hearing loss simulator 179

Figure 6.17 Boxplots showing the performance levels of the military and civilian samples on the CRM-CSoff (outliers exceed 1.5 times the interquartile range) 186

Figure 6.18 Boxplots showing the performance levels of the military and civilian samples on the VEHCOM SimMCAT (outliers exceed 1.5 times the interquartile range) 188

DECLARATION OF AUTHORSHIP

I, Hannah Domenica Semeraro, declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

DEVELOPING A MEASURE OF AUDITORY FITNESS FOR DUTY FOR MILITARY PERSONNEL

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published (in date order, journal publications in bold):

Study 1 part B: **Semeraro, H., Bevis, Z., Rowan, D., van Besouw, R. and Allsopp, A. 2015. Fit for the frontline? Identification of mission-critical hearing tasks (MCHTs) carried out by infantry and combat-support personnel. Noise and Health, 17(75), p. 98-107.**

Study 1 part A: Bevis, Z.L. and Semeraro, H. 2014. In brief: Fit for the frontline. University of Southampton New Boundaries Magazine, 19, p. 28.

Studies 2, 3 & 4: Semeraro, H., Rowan, D., van Besouw, R. and Allsopp, A. 2014, September. Developing a speech-in-noise test to measure auditory fitness for duty for military personnel: introducing using the Coordinate Response Measure. Aural presentation at the British Academy of Audiology Annual Meeting, Bournemouth, UK.

Studies 2, 3 & 4: Semeraro, H., Rowan, D., van Besouw, R. and Allsopp, A. 2014, September. Developing a speech-in-noise test to measure auditory fitness for duty for military personnel: introducing using the Coordinate Response Measure. Poster presented at the British Society of Audiology Annual Meeting, Keele, UK.

Study 1 part A: **Bevis, Z., Semeraro, H., van Besouw, R., Rowan, D., Lineton, B. and Allsopp, A. 2014. Fit for the frontline? A focus group exploration of auditory tasks carried out by infantry and combat support personnel. Noise and Health, 16(69), p. 127-135.**

Study 1 part B: Semeraro, H., Bevis, Z., Rowan, D., van Besouw, R. and Allsopp, A. 2014, April. Fit for the frontline? Identification of mission-critical hearing tasks (MCHTs) carried out by infantry and combat support personnel. Poster presented at the Biomedical Engineering, Science and Technology Research for Human Health Conference, Southampton, UK.

Study 1 part A: Semeraro, H., Bevis, Z., Rowan, D., van Besouw, R. and Allsopp, A. 2013, September. Fit for the frontline? A focus group exploration of auditory tasks carried out by infantrymen. Poster presented at the British Society of Audiology Annual Meeting, Keele, UK.

Signed: .....

Date:.....

Acknowledgements

Thank you to the Research Centre for Defence Medicine for funding the work completed in this PhD. I would like to thank all of the military personnel who have contributed towards the data collection and providing expert advice. I would also like to thank students at Southampton University who participated in experiments and anyone in Institute of Sound and Vibration Research who helped with experimental design and set up. The Defence Audiology Service was extremely helpful in facilitating data collection and supporting the design and running of Study 4, thank you Gerry Duffy and Pauline Tarrant.

I would like to thank my supervisor Dr Daniel Rowan for his support throughout my PhD. He went above and beyond the requirements of a PhD supervisor and provided me with invaluable life advice which has helped to mould my post PhD choices. Dr Adrian Allsopp supported me endlessly when completing MoD ethics applications and assisted me when facing the MODREC committee at Whitehall, London. Thank you Adrian for driving us all over the country to collect data, meet subject matter experts and observe training exercises.

Falk-Martin Hoffman your incredible MATLAB skills helped me get experiments up and running, thank you for teaching me a lot about writing code. Carla Perkins, thank you for giving up your evenings and weekends to patiently teach me MATLAB skills and help turn my pseudocode into an experimental procedure.

Zoë Bevis, we've been together through thick and thin! You've been an incredible friend, colleague, lunch and coffee buddy, listener, reality checker and advice giver, to say just a few. It has been invaluable to have you around throughout the whole PhD process. Also Matt Blyth, we were so thrilled when you joined the Hear for Duty Team, your endless enthusiasm is contagious and your constant willingness to help is amazing.

Finally, I'd like to thank my family for supporting me throughout the PhD process. My husband, Aaron Shutt has always been there for me and been incredibly understanding and patient.

Glossary

Below is a glossary for a selection of terms, which currently do not have a single definition within the literature. The definition provided represents how the term will be used within this thesis:

Mission-critical auditory tasks (MCATs) are military specific auditory tasks that are hearing dependent and failure to perform the task to a specified level will result in decreased safety and/or efficiency.

Hearing acuity refers specifically to an individual's ability to detect a sound, but not to their ability to make sense of the sound.

Hearing ability refers to an individual's ability to not only detect a sound and but also to make sense of it.

Auditory fitness for duty (AFFD) is the possession of sufficient hearing abilities for safe and effective job performance (Tufts et al, 2009).

Speech intelligibility refers to an individual hearing a speech signal accurately but does not including the comprehension of what has been said, e.g. measuring their ability to repeat what has been said but not a understanding of the speech.

Audibility refers to whether an individual is capable of detecting a sound, i.e. whether it is presented above their hearing threshold.

Measurement precision refers to the level of accuracy of a measurement tool. In this thesis, the measurement precision of the CRM as tool for assessing speech recognition thresholds is determined by the test-retest reliability and the concurrent validity of the test.

Abbreviations

AFFD	Auditory fitness for duty
BKB	Bamford-Kowel-Bamford
CATT	Combined Arms Tactical Trainer
CRM	Coordinate response measure
CRM-CSoff	Coordinate response measure- Call sign off scoring method
CRM-CSon	Coordinate response measure- Call sign on scoring method
dB A	Decibel A-weighted curve
dB HL	Decibel hearing level
dB SPL	Decibel sound pressure level
DCCT	Dismounted Close Combat Trainer
EASHW	The European Agency for Safety and Health at Work
GPT	Generic predictive test
H grade	Hearing grade
HCP	Hearing conservation programme
HCT	Hearing critical tasks
HI	Hearing impaired *(used in Study 4 only)*
HLS	Hearing loss simulator/simulation
ITDU	Infantry Trials and Development Unit
JSP	Joint Service Publication
MCAT	Mission-critical auditory tasks
MoD	Ministry of Defence
NH	Normal hearing *(used in Study 4 only)*
NIHL	Noise induced hearing loss
PF	Psychometric function
PRR	Personal Role Radio
PTA	Pure-tone audiometry
RMS	Route mean squared
SC-MCATs	Speech communication mission-critical auditory tasks
SIN	Speech-in-noise
SNR	Signal-to-noise ratio
SME	Subject matter expert
SRT	Speech recognition threshold
SRT 50	Speech recognition threshold at the 50% correct point
TDT	Triple digit test
TPT	Task-related predictive test
TST	Task simulation test

Chapter 1: Introduction

1.1 Background and motivation

The key motivation behind this thesis is to ensure that the occupational hearing standards used within the Armed Forces accurately predict whether personnel have adequate hearing, in order to carry out operational duties safely and effectively. There is a strong argument that any fitness for duty criteria should be based on evidence that shows whether individuals are able to carry out their job to a satisfactory standard and this is not currently the case within the UK military.

In a combat scenario, important information is often conveyed through acoustic cues, such as hearing commands over a radio or locating a weapon firing. Situational awareness is a term used to describe an individual being aware of what is happening around them and calculating the relative importance of different aspects of their surroundings (Stanton et al, 2001). A key aspect of situational awareness is 'information gathering', collating information about the surrounding environment from various sensory modalities (Endsley, 1995). Hearing loss may prevent an individual from detecting and utilising all the acoustic cues available to them, leading to reduced situational awareness.

Hearing loss is a particular problem for military personnel. Due to the nature of their work and the equipment they use, individuals are potentially exposed to damaging noise levels on a regular basis (NIOSH 2001-103, 2001). Regular exposure to high noise levels is known to cause noise induced hearing loss (NIHL; Grantham, 2012). A preliminary study, conducted by Surgeon Command Pearson, investigated the incidence of hearing loss during a tour of duty in Afghanistan (Operation Herrick 9). It was found that 42% of personnel had a measureable decline in hearing thresholds in comparison to their pre-deployment audiogram (Pearson, 2011). According to the Lost Voices report released by the British Legion, veterans under the age of 75 are three and a half times more likely to report hearing difficulty than the general population (The Royal British Legion, 2014). These statistics lead to the suggestion that hearing impairment, as a result of noise exposure, is particularly prevalent in the military population. In addition, personnel are susceptible to all other causes of hearing loss which affect the civilian population, such as presbycusis and conductive losses. Personnel must have sufficient hearing acuity in order to maintain situational awareness and carry out their operational duties safely and effectively; accurate hearing assessment within the Armed Forces is therefore considered a topic worth further investigation.

Chapter 1

Auditory fitness for duty (AFFD) refers to the possession of sufficient hearing abilities for safe and effective job performance (Tufts, 2011). Numerous occupations have AFFD protocols (such as the police force and fire fighters) and these all include auditory standards which employees must meet in order to continue with their job. However, the origin of these standards is often elusive (Tufts et al, 2009) and the AFFD protocol used by the Armed Forces is no exception. Pure-tone audiometry (PTA) is currently used in the UK to assess whether personnel have adequate AFFD. The military use results from PTA to decide whether personnel should be redeployed; their results are classified into one of five discrete groups, known as the hearing grades (H grades). PTA is a measure of tone detection in quiet and is known to be a good measure of audibility. However, it is widely accepted that suprathreshold psychoacoustic abilities (such as frequency selectivity and temporal resolution) are required when listening to more complex signals than the detection of a quiet pure-tone, such as speech (Summers et al, 2013). These suprathreshold abilities are not assessed by PTA, leading to the suggestion that PTA may not be able to accurately predict performance in more complicated listening scenarios, such as those experienced in a combat situation.

Measures of AFFD should be able to predict whether employees are able to complete workplace tasks that place demand on the individual's hearing, known as hearing critical tasks (HCTs; Laroche et al, 2003). For military personnel, these tasks have been termed mission-critical auditory tasks (MCATs; Semeraro et al; 2015). The MCATs identified by the author (Semeraro et al, 2015) all satisfy two characteristics; they are hearing dependent and failure to perform the task to a specified level results in decreased safety and/or efficiency. There are a total of 17 MCATs and they can be split into three broad categories of auditory skills: 1) seven *speech communication* tasks, *e.g.* accurately hearing grid references; 2) one *sound localisation* task, locating a small arms firing point; and 3) one *sound identification* task, identifying the type of weapon system being fired. The auditory skills required to perform these MCATs should be prioritised for representation in a measure of AFFD.

This thesis focuses specifically on ensuring that measures of AFFD are able to accurately predict performance on the speech communication MCATs (SC-MCATS). Pure-tone Audiometry (PTA) measures the audibility aspect of hearing impairment and not the additional processing deficits associated with sensorineural hearing impairment (Plomp, 1978). It is therefore not known whether PTA is able to accurately predict performance on the SC-MCATs or whether additional tests, such as a speech-in-noise (SIN) test, should be introduced as new measures of AFFD. This thesis explores the development of a SIN test, the Coordinate Response Measure (CRM), as a potential new tool for measuring AFFD, specifically focusing on the SC-MCATs.

1.2 Thesis structure, research objectives and dissemination of findings

Within this section an overview of each **CHAPTER** is provided, including details of the **research objectives** and the ***studies*** that are completed in each. This thesis has five studies and research objectives which are all working towards the ultimate goal of developing a tool for accurately predicting AFFD for military personnel. The publication and presentation of the author's work throughout the course of the PhD is detailed in italics. A full list of publications and presentations can be found in the declaration of authorship (publications) and Appendix N (presentations). For clarity, the research objectives are listed at the end of Section 1.2.

CHAPTER 2 provides an overview of the impact of hearing loss on situational awareness and explores the causes of hearing loss within the military, with a particular focus on NIHL. An introduction to AFFD and how it is currently assessed in the military is provided, outlining the potential problems with the current method. Considering that measures of AFFD within the military should be based on accurately predicting performance on MCATs, **CHAPTER 3** reports two studies that aim to identify the MCATs carried out by infantry and combat-support personnel (**research objective one**). This is achieved by using a series of focus groups to identify the auditory tasks carried out by infantry and combat-support personnel (***Study 1 part A***), and is followed by a questionnaire to gather information about which tasks satisfied the characteristics of a MCAT and should be prioritised for representation in a measure of AFFD (***Study 1 part B***). A decision was made to focus on this subset of the Armed Forces as the skills and tasks carried out by this population are the foundations of the initial training carried out by many other military groups, making it a logical starting point. This work has been published in Noise and Health Journal (*Bevis et al, 2014*; *Semeraro et al, 2015*) and presented at three conferences (B*ritish Society of Audiology Annual Conference [poster], 2013, Biomedical Engineering, Science and Technology Research for Human Health Conference [poster], 2014; British Academy of Audiology Annual Conference [aural presentation], 2014*) and at four seminars/workshops (*Institute of Naval Medicine Journal Club, 2013*; *Royal Centre for Defence Medicine, 2015*; *Imperial College Seminar Series, 2015; Imperial College Ear-Monitoring Workshop, 2015*).

Seven out of the nine MCATs are speech communication tasks (SC-MCATs). This thesis is focused specifically on the SC-MCATs and whether measures of AFFD are able to predict performance on these tasks. **CHAPTER 4** provides a theoretical argument as to why there is reason to suspect that PTA, the current measure of AFFD within the Armed Forces, may not be able to accurately predict performance on the SC-MCATs. A review of the literature shows that there is no clear agreement about the correlation strength between PTA and measures of speech intelligibility. There is reason

to suggest that a speech-noise-noise (SIN) test may be better able to predict performance on the SC-MCATs.

In **CHAPTER 5,** following a review of existing SIN tests, the Coordinate Response Measure (CRM) SIN test is chosen for further investigation, partly due to its high face validity when compared to command structure. The CRM sentence structure is "Ready *call sign,* go to *colour number* now". **Research objective two *(Study 2)*** is to design and record the British English version of the CRM. In order to implement the CRM test stimuli in an adaptive procedure (a method that allows for the rapid extraction of information about an individual's SIN ability; Leek, 2001), it is necessary to check that test stimuli meet certain criteria, including equal speech intelligibility across the speech corpus. **Research objective three** aimed to obtain speech intelligibility measurements for the individual call sign (colour and number target words of the CRM) presented in stationary speech-spectrum noise, and to adjust the stimuli amplitude, equalising the intelligibility of the CRM test material ***(Study 3)***. The final stage of developing the CRM adaptive procedure is to investigate the measurement precision of the CRM test adaptive procedure (**research objective 4**). In ***Study 4,*** the test-retest reliability of the CRM, with two scoring methods (responding to all three target words, 'call sign on', and responding to only the colours and numbers, 'call sign off') is investigated, as well as exploring the concurrent validity of the test in comparison to the Triple Digit Test (TDT), an alternative measure of SIN ability. For both scoring methods, the CRM adaptive procedure test has been shown to display adequate measurement precision to be considered a 'ready to use' SIN test. The design, development and evaluation of the CRM has been presented at two conferences (*British Society of Audiology Annual Conference [poster], 2014 and British Academy of Audiology Annual Conference [aural presentation], 2014*) and at two seminars (*Royal Centre for Defence Medicine, 2015 and Imperial College Seminar Series, 2015).* A journal publication of this work for the International Journal of Audiology (IJA) has been written and is currently being reviewed by the co-authors.

In order to evaluate and compare PTA and the CRM as tools for predicting performance on the SC-MCATs (their 'predictive validity') and ultimately for assessing AFFD, a method for measuring performance on the SC-MCATs is required. **CHAPTER 6** marks the final stage of this thesis and works towards developing a task simulation of the SC-MCATs. In order to measure the predictive validity of the CRM and PTA, as measures of AFFD, a test is required which is able to measure individual performance on the SC-MCATs. **Research objective five** was to design and develop a simulation for measuring performance on the auditory element of the SC-MCATs. ***Study 5*** addresses the design and development of a test that specifically focuses on the scenario of listening to commands over a radio in a moving vehicle. The simulation is named the Vehicle Communication Simulated MCAT, or VEHCOM SimMCAT. Prior to the VEHCOM SimMCAT being

used to assess the predictive validity of PTA and the CRM, there are still a number of unknown factors relating to performance on the simulation that need investigating. **Research objective 6 *(Study 6)*** is to evaluate whether performance on the VEHCOM SimMCAT is affected by hearing impairment and job experience, as would be expected for performance on the SC-MCATs. Study 6 concludes that the VEHCOM SimMCAT is sensitive to simulated hearing impairment and that performance on the task is affected by individual experience listening to military commands. The design and development of the VEHCOM SimMCAT and the findings of Study 6 have been presented at two seminars (*Royal Centre for Defence Medicine, 2015 and Imperial College Seminar Series, 2015)* and the author plans to publish this work in IJA. Further work is now required to develop a revised and improved SimMCAT and to use this tool to assess the predictive validity of PTA and the CRM, as tools for assessing AFFD. A proposal to carry out this work has been approved by the Royal Centre for Defence Medicine.

To summarise, a list of the research objectives achieved within this thesis are listed below:

Research objective one is addressed in Chapter 3, Study 1 part A and part B: to identify the auditory tasks carried out by infantry and combat support personnel, investigate which of these auditory tasks can be defined as MCATs and to decide which of the MCATs should be prioritised for representation in a measure of AFFD.

Research objective two is addressed in Chapter 5, Study 2: to design and record the British English version of the CRM.

Research objective three is addressed in Chapter 5, Study 3: to obtain speech intelligibility measurements for the individual call sign, colour and number target words of the CRM presented in stationary speech-spectrum noise and to adjust the stimuli amplitude to equalise the intelligibility of the CRM test material, so the necessary assumptions for implementation in an adaptive procedure are met.

Research objective four is addressed in Chapter 5, Study 4: to investigate and compare the measurement precision of the two CRM test adaptive procedure scoring methods and to investigate the concurrent validity of the test in comparison to the TDT, an alternative measure of SIN ability.

Research objective five is addressed in Chapter 6, Study 5: to design and develop a simulation for measuring performance on the auditory element of the SC-MCATs, focusing on the scenario of listening to commands over a radio in a moving vehicle.

Research objective six is addressed in Chapter 6, Study 6: to evaluate whether performance on the VEHCOM SimMCAT is affected by hearing impairment and job experience, as would be expected for performance on the SC-MCATs.

1.3 Contributions to knowledge

The six main contributions to knowledge from this thesis are:

1. A greater insight into the problems with the hearing assessment protocol currently used by the UK Armed Forces and, through extensive publication and attracting media attention, an increased awareness of the importance of accurate AFFD assessment, not only within the UK military but also in other occupations.

2. A methodological framework for identifying the HCTs carried out within any occupation which requires an AFFD protocol has been designed and developed. This methodology has been used to investigate the auditory tasks carried out by infantry and combat-support personnel and has produced a list of 17 MCATs carried out by this population. Nine of these MCATs are performed by the majority of ranks and roles either weekly or daily and have either major or critical consequence if performed poorly. These nine MCATs should be prioritised for representation by a measure of AFFD for infantry and combat-support personnel to ensure they have the necessary auditory skills for safe and effective deployment on operational duties.

3. The British English version of the CRM has been designed, recorded and evaluated. It has displayed adequate test-retest-reliability and concurrent validity, in comparison to SIN tests already used in audiology clinics, to be used as a tool for measuring speech recognition thresholds in stationary speech-spectrum noise. The re-recorded speech corpus was specifically designed to hold high face validity when compared to UK military command structure (using NATO call signs and UK military number pronunciation). The CRM test is not only ready to be investigated as a tool for assessing AFFD, but is also a useful tool for a variety of psychoacoustic experiments.

4. A simulation task has been designed and developed to measure performance on the auditory element of a subset of the SC-MCATs in a specific environment (listening to commands over a radio in a moving vehicle, the VEHCOM-SimMCAT). The initial development of this tool is a step towards being able to measure 'real world' performance, thereby allowing investigation of the predictive validity of potential AFFD tools.

5. A difference in performance levels on the VEHCOM-SimMCAT between military personnel and civilians was found. Military personnel consistently outperformed civilians when listening to commands that have been processed through a hearing loss simulator and are presented in adverse listening conditions. This finding suggests that military personnel are, to some extent, able to use their experience and knowledge of military communication to compensate for reduced hearing acuity. This finding leads to the suggestion that 'level of experience' and, potentially other top down processes, should influence personnel's AFFD.

6. The work in this thesis has provided a novel insight into the next steps that are required in order to move towards achieving the ultimate goal of developing a tool to accurately assess AFFD for military personnel. Further research proposals have been generated and acted upon (detailed in Section 7.3) as a direct result of this research. This further work will address topics raised in the Department for Health Action on Hearing Loss (The Department for Health, 2015) & the Lost Voices Report (The Royal British Legion, 2014).

Chapter 2: Auditory fitness for duty (AFFD) in the military

2.1 Introduction

The ability to listen to commands in noisy environments and understand acoustic signals, while maintaining situational awareness, is an important skill for military personnel and can be critical for mission success. Situational awareness describes an individual's awareness of what is happening around them and the relative importance of that information (Endsley, 1995). Due to the nature of their work and the equipment they use, military personnel are regularly exposed to unsafe levels of noise (NIOSH, 2001) and this puts them at high risk of NIHL. Personnel may also be affected by other common causes of hearing loss such as presbycusis, genetic hearing loss, and conductive hearing loss (caused by infection, build-up of wax, or a perforated ear drum). Situational awareness can be affected by one or more factors such as attention level, tiredness, stress, workload, experience, and, of particular interest here, impaired sensory modalities, such as hearing loss. One element of the information-gathering stage of situational awareness is picking up auditory cues. In a military operation, auditory information can be vital during the information-gathering stage of situational awareness. Not only is a great deal of information passed over radio communication systems but personnel also utilise environmental sounds to gain a detailed picture of their surroundings. This becomes particularly important when cues from other sensory modalities are obscured, for example when buildings or vehicles block the line of sight. In these situations the use of auditory cues to remain operationally effective is of utmost importance (Scharine et al, 2009). The phrase 'auditory fitness for duty' (AFFD) was first introduced by Tufts et al (2009) and refers to the possession of sufficient hearing abilities for safe and effective job performance. In the context of the Armed Forces, individuals with a hearing impairment may experience reduced situational awareness, thus impacting their AFFD.

2.2 Situational awareness

Situational awareness is a term used to describe a person being aware of what is happening around them and calculating the relative importance of different aspects of their surroundings. A commonly accepted definition of situational awareness is given by Endsley (1988, cited in Stanton et al, 2001), "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future". It is not difficult to imagine a scenario in which situational awareness is of utmost importance for

military personnel. For example, during a battle scenario a single serviceman may be attending to instructions over radio at the same time as operating a weapon system, whilst continuing to be fully aware of their surroundings and detecting the enemy's location.

Endsley (1995) proposes a theory that suggests three levels of situational awareness which have been summarised in Figure 2.1. Taking the three levels of situational awareness, three main components are identified: information gathering, perception of the information and making sense and giving meaning to these perceptions.

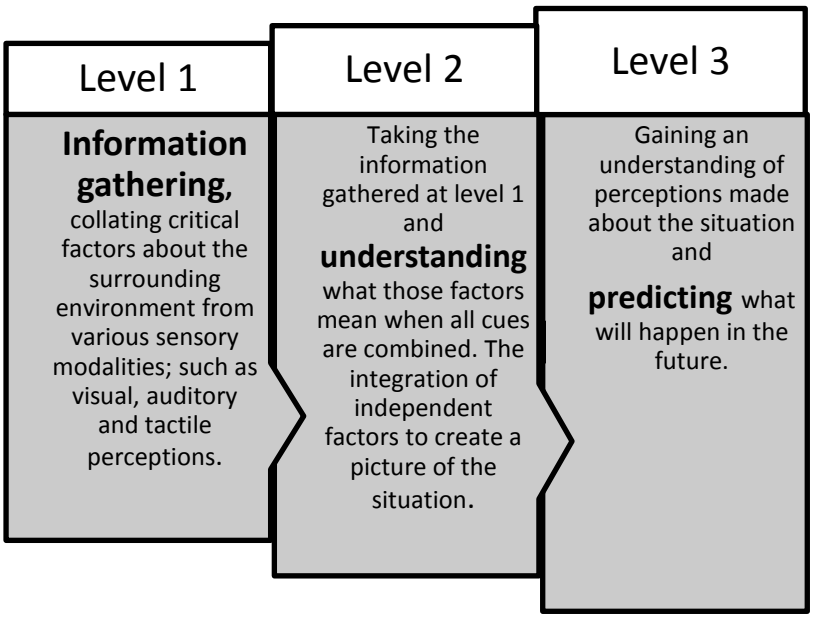


Figure 2.1 *The three levels of situational awareness outlined by Endsley (1995)*

Loss of situational awareness can be caused by one or a combination of factors such as tiredness, attention, stress, workload, job experience and impaired sensory modalities (e.g. hearing loss or impaired vision). Of particular interest in this project is the impact that hearing loss has on situational awareness. One element of the information gathering stage of situational awareness (Level 1) is picking up auditory cues. In a military operation gathering auditory information can be vital; not only is a great deal of information passed over radio communication systems, personnel are also required to utilise environmental sounds to gather a detailed picture of their surroundings. This becomes of particular importance when cues from other sensory modalities become obscured. For example, visual cues are often inaccessible because buildings or vehicles are in the way. In these situations the use of auditory cues becomes of utmost importance (Scharine et al, 2009).

Personnel who suffer from hearing loss will, as a result of reduced audibility, acquire less of the auditory signals available to them and therefore will not collect the maximum amount of information during Level 1 situational awareness (see Figure 2.1). There is an emphasis on

collecting as much data as possible about the surrounding environment; if important information is not gathered this will have implications for the higher levels of situational awareness as less information is being fed into the perception and understanding stages. For example, if the sound of a motorbike passing on a nearby road is not detected this may mean vital information about enemy movement has been missed. This influences whether an accurate picture of the current enemy location and predictions about the enemy's potential future actions can be made.

It is also important to note that the acoustic environment that military personnel experience in the battlefield is often not conducive to gaining accurate auditory information to feed into Level 1 situational awareness. Friendly and enemy fire can sound similar, poor radio connections are a possibility and continually competing background noise from vehicles, machinery and weapon systems mask important acoustic signals (Scharine et al, 2009). Additionally, some of the sounds being detected are much quieter and subtler then vehicles or weapons, such as footsteps and leaves rustling. These types of acoustic cues can however be vital if the enemy is fast approaching.

Military personnel need to maintain situational awareness in order to preserve operational effectiveness. Gathering all available information through all sensory modalities will ensure personnel maintain maximum situational awareness. However if personnel have a hearing impairment this may impact how much information is gathered during Level 1 situational awareness.

2.3 Hearing loss in the military

2.3.1 Introduction

It has been discussed in Section 2.2 that hearing loss can affect situational awareness, in particular during the information gathering stage. In occupations where workers are exposed to high levels of noise there are regulations to prevent employees suffering from NIHL. Due to the nature of their work and the equipment they use, military personnel regularly experience unsafe noise exposures (NIOSH, 2001). Exposure to high noise levels is arguably the main cause of hearing loss within the military. Service personnel are required to pass a hearing test prior to joining the Armed Forces so at this stage many causes of hearing loss which are present in the general population are detected and individuals are unable to join the military. It is therefore a reasonable assumption that the pre-enlistment rates of hearing impairment within the service population would be lower than the general population. However, according to the Lost Voices report released by the British Legion, veterans under the age of 75 are three and a half times more likely to report hearing difficulty than the general population (The Royal British Legion, 2014). This leads

to the prediction that hearing loss in the veteran population is probably more likely to have been caused by noise exposure.

Considering that hearing impairment is a greater problem in the military population than in the general population and given the importance of hearing acuity for maintaining situational awareness it therefore follows that hearing loss in the military is a topic worth further investigation.

2.3.2 Occupational noise exposure

The majority of adults spend most of their day in the work place. As employees they might face a number of hazards, depending on their type of work, potentially resulting in a range of effects on health. These health effects could be injury, stress or, of particular interest here, hearing loss. Hearing loss caused by work related noise is included in the World Health Organisation's list of major risks to health caused by occupational hazards (WHO, 2002). Noise in this context can be taken to be any sound that is unwanted and, in particular, noise that reduces the clarity of an acoustic signal. Exposure to loud sounds can cause a specific type of hearing loss known as noise induced hearing loss (NIHL) (Grantham, 2012). Sound protection limits have been set to protect workers from excessive noise exposure. The unit used to measure the amount of noise a worker is exposed to is given as the Equivalent Sound Level for the length of exposure time (Leq, h). Exposure time for a typical working day is 8 hours. The European Commission (EC) Directive has issued minimum health and safety requirements for the exposure of workers to risks arising from physical agents. Within the EC requirements are guidelines for noise exposure of Leq, 8h = 80 dB A (European Parliament, 2003). This regulation has been enforced since 2006 and contains three action levels for varying noise exposures over the 8 hour day and subsequent time periods, which are summarised in Table 2.1. These levels are based on the principle that a 3 dB A increase in sound level should result in halving the total exposure time, known as the time-intensity trade off.

Table 2.1 *European Commission (2008) guidelines on noise exposure regulations*

Exposure duration	Lower exposure action value (minimum)- protection must be provided (dB A)	Upper exposure action value- protection is mandatory (dB A)	Maximum exposure limit value (dB A)
8 hours	80	85	87
4 hours	83	88	90
2 hours	86	91	93
1 hour	89	94	96
30 mins	92	97	99
15 mins	95	100	102
1 min	107	111	113

2.3.3 Military noise exposure

The National Institute for Occupational Safety and Health (NIOSH, 2001), list the military as an industry in which workers are exposed to dangerous levels of noise and at risk of NIHL. Military personnel are exposed to two types of sounds; continuous and impulse (Cain, 1998). Continuous noise is defined as noise in which the highest levels occur more than once per second. Examples include helicopter crew exposed to engine noise of to up to 102 dB(A) and fighting vehicle drivers, such as tanks, which can have internal noise levels of 115 dB(A) (Cain, 1998). Impulse noise can be used to describe all high-level and short-duration sounds. This incorporates a range of sounds with varying durations (microseconds to milliseconds), intensities (less than 100 dB sound pressure level (SPL) to over 185 dB SPL) and spectral properties (impulse noise often covers a broad range of frequencies but some noise may contain peaks at particular frequencies). A common example of impulse noise within the military would be small arms (Henderson & Hamernik, 1986).

The noises to which any one person in the military is exposed will vary and are dependent on a number of factors such as his/her job role and the activities he/she performs. It is thought that the most harmful noise to which military personnel are exposed is the impulse noise from firearms (Ylikoski & Ylikoski, 1994). There is also a risk of hazardous noise levels from continuous engine noise and communication systems (Muhr, 2010). Explosions and other weaponry can also generate damaging noise levels.

The specific noise types to which military personnel are exposed are dependent on a number of factors (e.g. weapon system or vehicle type). Continuous noise is commonly generated by vehicle noise. There are numerous factors which cause variation in the level of noise a vehicles will generate: vehicle type; engine system; weapons used from within the vehicle; communication equipment; terrain; speed; driver skill, experience and style; position within the vehicle; loading; and state of doors, hatches and windows (NATO, 2010). To give some idea of the noise exposures, two examples are given below.

1. Personnel travelling in the cargo compartment of the Super Hercules military transport aircraft can experience noise exposures of 118 dB(A) (NATO, 2010). In the cockpit of a Chinook and Apache helicopter noise exposure of 103 and 101 dB(A) are heard, respectively (Humes et al, 2006). For land based vehicles, such as the UK Warrior and Challenger heavy tanks, the noise exposures can be 110-115 dB(A) when driven in a worse case noise condition scenario, such as along a hard surface at high speeds (NATO, 2010).

2. Impulse noise is generated by weapon systems and also produces high noise levels. For example, small calibre weapons, such as hand guns and assault rifles have noise exposures of roughly 160 dB(A) when recorded at the position of the shooter (NATO, 2010) and a 9mm pistol generates 157 dB (A) (Humes et al, 2006). Larger weapons, such as howitzers or mortars generate noise exposures as high as 190 dB(A); a 105mm towed Howitzer generates 183 dB(A) at the position of the gunner (Humes et al, 2006).

It is clear from these values that the sound levels experienced by military personnel are above the recommended exposure levels outlined by the European Commission (2008; see Table 2.1). If military personnel are routinely exposed to noise levels similar to those outlined above and they are not consistently wearing hearing protection then this will put these individuals at risk of NIHL.

2.3.4 Noise induced hearing loss

Noise induced hearing loss (NIHL) is hearing loss caused by exposure to loud sound, defined as over 80 dB A by the European Parliament (Grantham, 2012). This can be either a single loud sound or continuous exposure over an extended period of time. Section 2.3.4 discusses the issues surrounding NIHL, including an overview of the pathophysiological effects of noise and the diagnosis of NIHL. Exposure to high levels of noise results in sensorineural hearing loss; NIHL is an entirely preventable form of acquired sensorineural hearing loss. Acoustic overexposure causes structural damage to the cochlear and has an effect on the neural coding of sound, the degree of which depends on individual susceptibility. This damage results in raised hearing thresholds and therefore reduced hearing acuity. The extent and origin of the damage caused is affected by the length and sound level of exposure but also by the characteristics of the noise such as the frequency and whether it is steady, fluctuating or impulse (Hu, 2012).

Exposure to high level sounds has been found to cause progressive deterioration of the auditory system. Noise can cause both direct mechanical damage to the middle ear and damage to the delicate structures of the inner ear, such as the sensory cells, nerve endings and the vascular supply. An audiogram typical of an individual exposed to excessive noise may contain a notch in the air-conduction thresholds at 4 kHz, where the hearing threshold levels at 3 and/or 4 and/or 6 kHz are at least 10 dB hearing loss greater than at 1 or 2 kHz and at 6 or 8 kHz (Coles et al, 2000).

The following paragraphs explore the pathophysiological effects of noise exposure when travelling through the hearing pathway. Firstly we will address how this characteristic audiogram feature of NIHL can be explained by features of the middle ear (ME). The external auditory meatus (EAM) has a resonant frequency, on average, of 3200 Hz. At this point of acoustic resonance the sound pressure level can increase by as much as 20 dB SPL. Vibrations along the basilar membrane have

been shown to have maximum displacement at half an octave above the stimulated frequency region. The causes more damage to the cochlea in this region. Therefore, the transform of the middle ear coupled with the half octave shift of fundamental EAM resonance causes the 4 kHz notch, typical of NIHL.

There is also a middle ear function which plays an important role in protecting the ear from acoustic overstimulation, known as the acoustic reflex. Contraction of the stapedius muscle and the tensor tympani muscle stiffens the ME, which decreases the transmission of sound through the middle ear. This attenuating action occurs only for sounds below 2 kHz (Borg et al, 1982). Individual variation in strength of acoustic reflex will affect responses to loud noise. An example of this is shown in Bell's Palsy sufferers who, when exposed to bilateral acoustic overexposure, developed substantially more of a temporary threshold shift on the Bell's palsy side compared to their normal side (Zakrisson et al, 1980, cited in Borg et al, 1982).

Moving onto the inner ear, the most significant damage caused by acoustic overexposure is the mechanical stress put on the cochlear structures. Loud sounds cause the basilar membrane to move excessively, in turn causing structural damage to cochlear sensory cells and their supporting cells and ultimately reducing cochlear function (Hu, 2012). According to Hu (2012) there are two basic causes of damage. Firstly, the direct mechanical stress, resulting from the physical impact of the sound, which can be detected immediately. Secondly, metabolic disruption can progress days or weeks after exposure. This metabolic disruption can be caused by a number of pathological conditions such as ischemia (restriction of blood flow), excitotoxic damage (damaged caused by excessive stimulation by neurotransmitters) or the intermixing of cochlear fluids. Acute high- level acoustic trauma can cause wide spread fractures of the tight cell junctions of the Organ of Corti as well as entire sections of cell separating from the basilar membrane.

Greater damage is seen in outer hair cells (OHC) compared to inner hair cells (IHC); this is thought to be because of three reasons. Firstly, the OHC are subject to direct force from the acoustic stimuli whereas IHC are only stimulated by the viscous drag. Secondly, the OHC are unsupported in comparison to the IHC which have supporting cells. Thirdly, the OHC are closer to the point of maximum basilar membrane travelling wave displacement in comparison to IHCs (Henderson & Hamernik, 1995).

The type, intensity and duration of the noise have a direct impact of the severity of the hair cell lesion it causes. Longer durations of noise exposure cause greater damage than shorter durations of the same intensity (Erlandsson et al, 1980). In terms of the effect of intensity, there is not a direct correlation between an increase in SPL and the level of hair cell damage caused. It is thought there is a critical level, below which the damage caused increases slowly and above which

the number of cell lesions increases significantly (Hu, 2012). It is not fully understood what causes this sudden increase in cochlear damage but it is thought that potentially, up until the noise reaches the critical level, the damage caused is metabolic but, above the critical level the damage shifts to become mechanical (Hu et al, 2000; Yang et al, 2004). The amount of damage is also affected by the frequency of the noise. It has been shown that high frequency noise causes more damage than low frequency noise (Erlandsson et al, 1980).

At the very early stages of exposure, high noise levels have a damaging effect on the synapses of the OHC and IHC. High noise levels cause signs of high levels of metabolic activity in IHC and OHC, such as severe swelling, which leads to swelling and death of dendritic terminals (Spoendlin, 1971). Overexposure has been shown to cause acute loss of afferent nerve terminals and overtime degeneration of the eighth nerve. It is thought that some of this degeneration may regenerate over time, possibly explaining some of the recovery seen from temporary threshold shifts (Kujawa & Liberman, 2009).

Each of the changes to the hearing system described above will result in a combined effect on hearing threshold levels. Changes to the cochlear structure are reversible or irreversible, depending on the amount of damage. Changes to the IHC and OHC can cause changes in threshold as well as the tuning characteristics of the eighth nerve fibre. Studies have shown that stereocilia damage may not directly cause hearing damage in terms of audibility, since substantial hair cell loss has been reported in individuals who still present audiometric results within the normal range (Henderson & Hamernik, 1995). There is also within-species variation in the amount of damage caused when individuals are exposed to the same noise. Studies with chinchillas found that, when exposed to a 161 dB SPL impulse noise, animals with the same threshold before exposure displayed different amounts of hearing loss. The range of variability seen is similar to that seen in human demographic studies (Henderson & Hamernik, 1995).

In summary, acoustic overstimulation can cause pathological damage at numerous points along the hearing pathway. Although a high frequency notch in the air conduction measurements with audiometry is indicative of noise damage it is not possible to assume a link between audiometric thresholds and the site of pathological damage. It is also important to remember that the effect of noise exposure can vary between two individuals who have been exposed to the same noise. It is worth noting that the majority of the studies referred to in this section are carried out using animals (Erlandsson et al, 1980; Yang et al, 2004; Spoendlin, 1971; Kujawa & Liberman, 2009), presenting challenges about the generalisability of the results to humans.

2.3.5 Additional causes of military hearing loss

Military personnel are of course susceptible to all the common causes of hearing loss, such as presbycusis, genetic hearing loss and conductive hearing loss. Although these causes are not addressed in detail here it is important to acknowledge they will be prevalent within military populations. There are however two causes of hearing loss, in addition to noise exposure, which are of particular interest for military personnel: conductive hearing loss caused by blast injury and the synergistic damage caused by ototoxicants and noise exposure.

When in the battlefield, military personnel may be in close contact with high charge explosives. When these explosives are detonated a pressure wave is generated that is powerful enough to injure the personnel exposed to them. The damage that this causes is known as barotrauma which occurs when the pressure between internal organs and the outer surface of the body differs at the moment of pressure wave impact (Argyros, 1997). The ear is an air-filled organ and therefore structures of the ear may be affected by barotrauma. Most commonly, the tympanic membrane will perforate as a result in the sudden change in pressure between the outer body and the ME. This can cause a 0-30 dB hearing loss in the low frequencies. Middle ear injury may also result in fractured ossicles and/or displacement of the stapes resulting in a conductive hearing loss of 0-25 dB HL (Mayorga, 1997). In addition to directly damaging the middle ear individuals may experience blast-related brain injury (Kocsis & Tessler, 2009). Depending on which areas of the brain are damaged this may result in difficulties processing sounds.

Ototoxicity refers to damage caused to the ear by a toxin. Ototoxicants refers to all substances that have an effect on the structures and/or function of the inner ear (cochlea and vestibular system). Ototoxicants can be split into cochleotoxicants and vestibulotoxicants. Cochleotoxicants impair cochlear structures including the hair cells, stria vascularis and spiral ganglion cells. Vestibulotoxicants cause damage within the vestibular system (The European Agency for Safety and Health at Work (EASHW), 2009). There are more than 200 ototoxic medications and chemicals known to cause hearing and balance problems (Cone et al, 2013). Although it is not necessarily the case that military personnel are more likely to take ototoxic drugs as part of their occupation there is a synergistic effect when an individual is exposed to ototoxicants and exposed to noise simultaneously. “Experiments with rats have shown that combined exposure to noise and certain solvents induced synergistic adverse effects on hearing” (The EASHW, 2009 pp. 28). Studies have reported that the protective role of the acoustic reflex is reduced when an individual is consuming ototoxic medication (The EASHW, 2009). The combined damage of ototoxicants and noise exposure is of particular relevance to military personnel who are exposed to particular high noise levels (see Section 2.3.3) and therefore could be more susceptible to NIHL

when taking ototoxic medication. Following a period of taking such medication it would be necessary to assess whether any hearing damage has been caused.

2.3.6 Summary

As part of their work military personnel are exposed to high levels of noise, putting them at increased risk of NIHL. Personnel are also affected by other more common causes of hearing loss. Within Section 2.2 the impact hearing loss may have on the information gathering stage of situational awareness was discussed. This leads to questioning how the hearing ability of military personnel is assessed and whether this assessment process is capable of measuring the impact hearing loss has on situational awareness. Section 2.4 addresses these issues, introducing of the concept of AFFD.

2.4 Auditory fitness for duty

2.4.1 Introduction

Auditory fitness for duty refers to the possession of sufficient hearing abilities for safe and effective job performance (Tufts, 2011) and is important in occupations in which employees are relying on acoustic cues in order to maintain situational awareness and perform their duties. These tasks or activities must be undertaken effectively and efficiently to a standard which does not compromise either their own safety or that of others. A number of occupations have AFFD protocols, including driving public transport, firefighting, law enforcement, manufacturing and, of interest here, the military (Tufts et al, 2009). One common factor amongst these professions is that they all require individuals to maintain situational awareness in order to carry out the job safely, effectively and efficiently. As discussed in Section 2.2, one element of situational awareness is being able to gather acoustic cues. For each profession, there are auditory standards which employees must meet in order to continue with their job. Although there are many occupations which have AFFD protocols the origin of these standards is often elusive (Tufts et al, 2009). There is however a common starting point when considering AFFD testing; determining which workplace tasks place demand on the individuals hearing ability. These tasks are termed hearing critical tasks (HCTs).

2.4.2 Hearing critical tasks

Within the literature discussing AFFD there are two commonly quoted definitions of HCTs. Firstly, Laroche et al (2003) defines a task as hearing critical only if it can be performed to a specified

level of accuracy by a normal hearing person using the sense of hearing alone (for the purposes of this report normal hearing is defined as average hearing thresholds at 250, 500, 1000, 2000 and 4000 Hz of 20 dB hearing loss or lower (British Society of Audiology, 2012). This definition is focused on performance level and regards a task as hearing critical if a reduction in performance is observed by individuals with impaired hearing. Tuft et al (2009 pp. 546) modified this definition and considered the consequences of poor performance, taking HCTs to include "tasks for which hearing loss would be a liability in inexperienced workers". To satisfy this definition there must be negative consequence(s) of some description caused when a task is carried out below a specified level.

For this report an alternative definition is proposed which combines the definitions from Laroche et al (2003) and Tuft et al (2009) and considers hearing dependency, performance level and consequence of poor performance; specifically addressing the critical component of HCTs. If critical is taken to mean "having the potential to become disastrous; at a point of crisis" (Oxford English Dictionary, 2015), then it follows that the potential consequences of poor performance of a hearing dependent task must be explored in order to confirm whether they will be adverse. Consequently, for the purposes of this report, in order for a task to be deemed hearing critical it must satisfy the following two characteristics:

1. The task must be hearing dependent
2. Failure to perform the task to a specified level must result in decreased safety and/or efficiency.

Regarding the first characteristic, a hearing dependent task can be performed to a specified standard by a normal hearing worker and performance will decrease when a hearing impaired person carries out the task. For the second characteristic, the threshold between poor and good performance is determined by identifying the point at which performance level is compromising the safety of the worker and their colleagues, as well as reducing efficiency and/or effectiveness; this is commonly determined by subject-matter experts (Giguère et al, 2008; Laroche et al, 2008; Laroche et al, 2003; Tufts et al, 2009; Forshaw et al, 1999).

It should be emphasised that the above definition remains a proposal and that no general consensus exists within the literature regarding the definition of a hearing critical task. There are however additional factors that the proposed definition does not take into account, such as the influence of multimodal integration on performance for HCTs. For example, if an individual is able to perform the task to a certain standard using other sensory modalities, such as sight or touch, then this may have an impact on whether a task is entirely hearing dependent. Furthermore, hearing dependency is not necessarily binary; it may not always be the case that those with

normal hearing will perform the task above a specified level 100% of the time and those with abnormal hearing will display decreased performance 100% of the time. It could be that an individual with a hearing impairment is more likely to display decreased performance on a task but may also carry out the task with no effect on performance a certain percentage of the time. This is important to consider when assessing whether a task is hearing dependent or not.

This particular project is looking at AFFD testing within the military. The HCTs for military personnel are termed 'mission-critical auditory tasks' (MCATs). These tasks must still possess both the characteristics of a hearing critical task but naming them 'mission critical' highlights the fact that it is the military's duty to carry out missions with maximum safety and operational effectiveness and efficiency (also taking lethality into account).

2.4.3 Auditory fitness for duty assessment and challenges

Measures of AFFD should be based on measuring whether employees are capable of carrying out job specific HCTs safely and effectively. For the majority of occupations the assessment of AFFD almost always involves testing audiometric thresholds with pass/fail cut off values for each frequency tested (Tufts et al, 2009). This is a simple way of testing AFFD; if the individual has normal hearing thresholds (as defined in occupational standards for the given job) then it is assumed they are capable of performing their job. Using pure-tone audiometry (PTA) alone to measure AFFD assumes a relationship between hearing acuity and job specific HCTs. An assessment of the predictive validity of PTA as a measure of AFFD is given in Chapter 4. Section 2.4.3 focuses on how AFFD is currently assessed in a generic sense and Section 2.5 explores the current methods for assessing AFFD within the Armed Forces.

The Equality Act (2010) explains that any exclusionary criteria for employment must be related to the job in hand and be connected with individual ability to perform the job safely and effectively. Any exclusionary criteria which do not satisfy the Act would be viewed as discriminatory. This is an important consideration when designing an AFFD test for most occupations, although some professions are exempt from these rules (military personnel with front line duties being an example). The Equality Act (2010 pp. 142) talks about 'relevant discrimination' which incorporates allowing discrimination against anyone who could potentially reduce the operational effectiveness of the Armed Forces. This includes allowing discrimination against age, disability (including hearing impairment), gender reassignment and sex. For an AFFD measure to not be considered discriminatory there must a clear relationship between the test results and job performance.

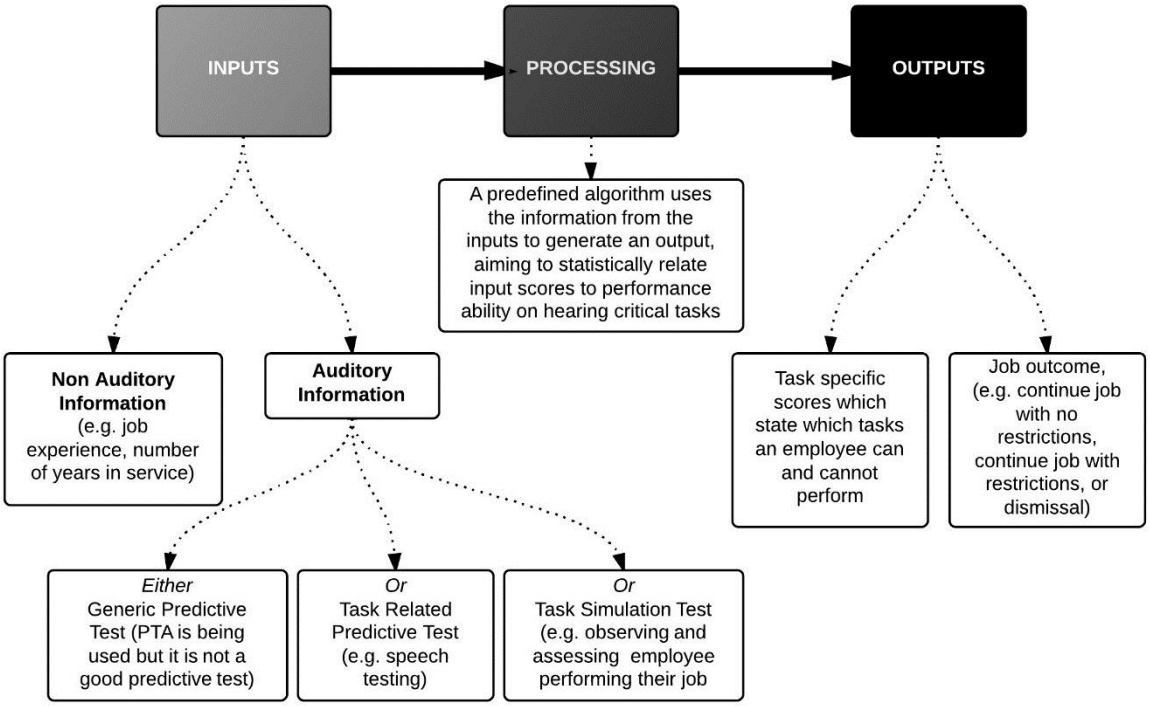


Figure 2.2 *Flow diagram showing generic AFFD testing methods*

A simplified way to think about AFFD tests is to consider the inputs and outputs, as shown in Figure 2.2. Firstly, the inputs will be addressed. Non-auditory information, such as job experience or number of years in service, may have an impact on performance on HCTs. An experienced worker who has gradually lost some of their hearing may still be able to function safely and effectively in situations where an inexperienced worker might struggle. An experienced worker could compensate for their hearing loss by relying on skills and knowledge gained from years of experience (Tufts et al, 2009). Consequently, it may, in some occupations, be necessary to take these factors into account when testing AFFD.

The auditory inputs include any auditory tests performed as part of AFFD assessment. When developing a physical fitness for duty test, Payne and Harvey (2010) divided the input test type into three categories; generic predictive tests (GPTs), task related predictive tests (TPTs) and task simulation tests (TSTs). The same approach is used here. These three categories should not be thought of as discrete groups but rather a continuum in which GPTs and TSTs are at opposite ends and TPTs are a type of hybrid test with characteristics of the GPTs and TSTs (Payne & Harvey, 2010). Generic predictive tests have no specific job related characteristics but can be used to predict performance on HCTs. It may initially be thought that PTA could be included in this test type but the ability of PTA to predict performance on HCTs cannot be assumed (this is discussed further in Chapter 4). At the other end of the spectrum are TSTs which are based directly on the HCTs being performed; the actual job characteristics will be preserved as far as possible for a TST. This may involve observing the individual performing their job and scoring their performance

accordingly and may be based on one specific or several different HCTs. A TPT is not based on any specific hearing critical task but is a test which is able to predict performance for a number of tasks. An example of this may be speech tests which are more similar to 'real world' listening situations than a GPT. As is the case for GPT, the results from a TPT must be statistically known to be able to predict performance to a known level of accuracy for job specific HCTs.

The next stage of an AFFD test is the processing. This part of an AFFD will vary depending on the input and the desired output. Algorithms should be developed having performed experimental work looking at the predictive ability of the input test (e.g. GPT or TPT) for the desired output. For example, if the input is a speech test then the speech test score must be statistically known to predict (to a known level of accuracy) performance on HCTs.

Finally, the output(s) of the test is produced. The most important use of an AFFD test is to decide whether or not an individual is able to continue to perform their job or whether their duties should be restricted or removed altogether. If a TST approach has been used then the output may be able to specifically state which HCTs an individual is capable or incapable of performing to the required standard. If a GPT or TPT has been used then this is more difficult to ascertain unless experimental work has been carried out looking at the relationship between performance on the AFFD test and performance on HCTs.

The process described in Figure 2.2 allows for a simplified overview of AFFD testing to be given but the model is dependent on a number of factors which should not be overlooked when considering the difficulty of AFFD assessment.

1. Identifying HCTs. Without an understanding of the HCTs within a specified occupation it is not possible to know which auditory skills are required to carry out the job. Identifying these HCTs is not always a simple or quick procedure.
2. Predicting job performance. Selecting a test which is able to predict performance on the HCTs, to a known level of accuracy, is not simple. Developing methods for assessing performance on HCTs which reflect occupational performance is not a simple task. In addition, there is always a certain amount of error when using one performance measure to predict performance on another.
3. Quantifying non-auditory inputs. A process for quantifying non-auditory inputs is not a simple requirement. For example, individuals may use other sensory modalities to carry out HCTs, such as sight, but this could vary greatly between individuals and measuring it may not be simple.
4. Sensitivity and specificity of AFFD test. The model relies on the selected AFFD measurement tool being capable of distinguishing between those who are unfit for duty

and those who are still capable of carrying out their job. Selecting a 'cut-off' point with high sensitivity and specificity may be difficult.

2.4.4 Summary

It has been established that measures of AFFD should be based on predicting performance on job specific HCTs. An overview of a generic measure of AFFD has been given in Figure 2.2 and some of the difficulties associated developing these tests have been discussed at the end of Section 2.4.3. Section 2.5 will explain how AFFD is currently assessed within the Armed Forces. Towards the end of Section 2.5 the suitability of the current measure will be addressed and the justification for questioning the current methods for assessing AFFD within the Ministry of Defence (MoD) will be explained.

2.5 Military auditory fitness for duty assessment

2.5.1 Introduction

The MoD operates an AFFD protocol based on pure-tone audiometry (PTA). This protocol has been developed as part of a hearing conservation programme (HCP), which aims to identify those who are most susceptible to NIHL and to ultimately prevent NIHL within the Armed Forces. The most recent documentation with regards to assessing AFFD and implementing the HCP is the Joint Service Publication 950 (JSP 950; MoD, 2013). This document outlines the recommendations for regular hearing screening using PTA and suggested actions for personnel with various levels of hearing acuity.

This assessment tool could be considered a GPT (see Figure 2.2) since it does not directly measure performance on any job specific HCTs but the results are being used to predict job performance. It should be noted however that the suitability of PTA as a tool for predicting performance on MCATs is currently unknown; this is further discussed in Chapter 4.

The assessment of hearing is part of the PULHHEEMS system which is used for grading the physical and mental fitness of Britain's Armed Forces (Biggs & Everest, 2011). PULHHEEMs is an abbreviation for: Physique; Upper limbs; Lowering limbs (Or 'Locomotion, as this includes the back); Hearing left; Hearing right; Eyesight left (corrected/uncorrected); Eyesight right (corrected/uncorrected); Mental function; Stability (emotional). The PULHHEEMS is designed to assess a serviceman's fitness for duty, giving a PULHHEEMS Employability Status (PES) which determines the areas in which the solider is able to serve. Each factor within PULHEEMS has its own method of scoring which ultimately results in an overall PULHEEMS score from 1 to 8 (1 being

excellent and 8 being unfit for service). For the factor 'Hearing', the current scoring system is contained in five distinct groups known as the H grades (H1, 2, 3, 4 and 8). Section 2.5.2 outlines the hearing assessment protocol for the Armed Forces and how the H grade system works.

2.5.2 Armed Forces hearing assessment protocol

The JSP 950 outlines two types of PTA: screening and clinical. It is documented that both of these procedures should be carried out using the procedure outlined in the British Society of Audiology recommended procedure for PTA (British Society of Audiology, 2012). However for the screening procedure, automated PTA is recommended over the manual method. It is stated that the clinical PTA should be carried out only by trained personnel, such as audiologists. According to the JSP 950 the regular scheduling of PTA aims to detect NIHL at an early stage and to allow for precautionary measures to be put in place if an individual is considered at a high risk of NIHL. All service personnel have pre-employment PTA carried out and following this an annual screening PTA is carried out during their first two years of employment. This aims to identify individuals that acquire hearing loss during basic training and personnel that may have an increased propensity to NIHL. After this point the regularity of PTA screening is dependent on: 1) whether the individual works in a noise hazardous area; 2) if the individual has acquired a hearing loss during basic training, suggesting they may have an increased propensity to NIHL; 3) if the individual is completing pre/post deployment medical checks (MoD, 2013). However, regardless of these factors all personnel (excluding civil servants) should have their hearing checked a minimum of every two years.

The MoD categorise all PTA results into hearing acuity grades, commonly referred to "H grades". These grades are calculated for each ear separately and are based on the sum of the low frequency (0.5, 1 and 2 kHz) and high frequency (3, 4 and 6 kHz) pure-tone thresholds. The highest value sum, either the low or high frequency sum, is taken to determine which H grade is assigned for each ear. Results are usually shown as HRightHLeft. The individual is then categorised based on the H grade of their worst ear. There are five H grades: 1, 2, 3, 4 and 8, which are outlined in Table 2.2.

The JSP 950 outlines that PTA results and the calculated H grade should be used for assessing the following four key areas discussed below (outlined in JSP 950; MoD, 2013). The outcome of this assessment is then reported as either HCP pass or HCP referral, the result of which is given to the individual immediately.

1. Assessment of fitness for role. This can be considered as a basic measure of AFFD. H grades are compared to the individual's single service guidance, to ensure their H grade

allows them to continue with their current employment. The single service guidance provides role-specific information about the physical and mental requirements for each role but these documents are not publically available. If the H grading has not changed in comparison to previous assessments then the individual is deemed still fit for duty (HCP pass). If a change from H1 to H2 is identified the individual should be referred to Occupational Health (HCP fail). If a change from H1/2 occurs the individual should be referred to the Defence Audiology Service (HCP fail). The individual's new H grade will be compared to the single service guidance, enabling a decision to be made regarding whether they can continue in their current role.

2. Assessment of rapid hearing loss. A rapid loss is defined by JSP 950 as a loss of 30 dB hearing level (HL) or more in the high frequency summed pure-tone threshold in the last three years. Providing that previous audiograms exist and have been recorded this can be calculated. If no such change is identified the individual continues with routine screening PTA (HCP pass). If such a change is observed the individual should be referred to the Defence Audiology Service (HCP fail).
3. Assessment of unilateral hearing loss. Asymmetric hearing is defined as a difference in the sum of the pure-tone thresholds at 1, 2, 3, 4 and 6 kHz for each ear of 45 dB HL or more. If an asymmetry is not identified the individual continues with routine screening (HCP pass). If an asymmetry is detected the individual should be referred to the Defence Audiology Service (HCP fail).
4. Assessment based on age and gender. A table of age and gender specific hearing requirements is provided in the JSP 950 (shown in Table 2.3) which provides 'warning' and 'referral' hearing acuity levels based on summed pure-tone thresholds for 1, 2, 3, 4 and 6 kHz. For those below the warning level no action is required (HCP pass). For those within the warning level the individual and their line manager are informed and the individual is required to attend annual hearing screens (HCP fail). If an individual has hearing thresholds within the referral category they are referred to the Defence Audiology Service (HCP fail). It is worth noting that the JSP 950 does not give reference as to where the data in Table 2.3 is derived from.

Table 2.2 *Armed Forces hearing acuity grades (MoD, 2013)*

Grades	Sum of dB HL at low frequencies in dB HL	Sum of dB HL at high frequencies in dB HL	General Description
1	Not more than 45 (no single level to be more than 20)	Not more than 45 (level not to be more than 30 at 6 kHz, or 20 at any other frequency)	Good hearing
2	Not more than 84	Not more than 123	Acceptable hearing
3	Not more than 150	Not more than 210	Impaired hearing
4	More than 150	More than 210	Poor hearing where continuing employment is subject to specialist assessment
8	More than 150	More than 210	Poor hearing that has been assessed as being incompatible with continued service

Table 2.3 *Age and gender specific hearing threshold assessment (MoD, 2013)*

Sum of dB HL at 1, 2, 3,4 and 6 kHz				
Age (years)	**Males (dB)**		**Females (dB)**	
	Warning	**Referral**	**Warning**	**Referral**
18-24	51	95	46	78
25-29	67	113	55	91
30-34	82	132	63	105
35-39	100	154	71	119
40-44	121	183	80	134
45-49	142	211	93	153
50-54	165	240	111	176
55-59	190	269	131	204
60-64	217	296	157	235

2.5.3 Summary

Currently the MoD has a HCP in place that is primarily designed to detect NIHL at an early stage. Parts of the HCP also aim to address whether an individual is fit for duty, such as the provision of single service guidance and the regular monitoring of hearing acuity levels, especially pre-deployment. However the current HCP cannot be said to adequately measure AFFD for two main reasons.

Firstly, the H grade cut offs given in the single service guidance documents are not based on evidence that individuals with the specified hearing requirements are fit for duty, that is to say they are capable of carrying out the MCATs for the role in question. Furthermore, a literature search by the author and staff at the Institute of Naval Medicine has not unveiled any evidence which justifies the cut off points between the five H grades. If it is in fact the case that they are arbitrary grouping then this raises additional concern about their use as fitness for duty standards. Secondly, the tool which is currently being used to measure AFFD is PTA; any measure of AFFD should be based on assessing performance on the job specific HCTs. This is problematic for two reasons: 1) there is currently no record of the MCATs carried out by military personnel (this is addressed in Chapter 3) and 2) given that PTA tests an individual's ability to detect a pure-tone in quiet there is reason to question whether the audiogram can be used to predict performance in more complex listening environments (addressed in Chapter 4).

2.6 Chapter 2 summary

Auditory fitness for duty testing is necessary whenever people are carrying out HCTs as part of their occupation; to develop an AFFD test it is first necessary to identify what these tasks are. It is then important to ensure that any AFFD test has predictive validity; that is to say the test is statistically known to predict ability to perform HCTs. In a review of AFFD tests, conducted by Tufts et al (2009), it was found that there is often little or no explanation of the relationship between the chosen AFFD test (usually based on audiometric thresholds) and safe and effective job performance. The military is no exception; there is no evidence to show the origin of the H grades and, furthermore, there has been no research looking at the predictive validity of PTA for predicting performance on MCATs and ultimately assessing AFFD. A gap in knowledge has been identified in this area; no catalogue of sounds to which military personnel are exposed is available (Grantham, 2012) and furthermore, no work has been done regarding the mission criticality of hearing dependent tasks carried out by the military. Chapter 3 outlines a study conducted to identify the MCATs carried out by infantry and combat support personnel, the first stage towards a developing a measure of AFFD.

Chapter 3: Identification of mission-critical auditory tasks (MCATs)

3.1 Introduction

Measures of AFFD should be based on job-specific hearing critical tasks (HCTs). In a military context these HCTs have been termed mission-critical auditory tasks (MCATs). For a task to be deemed a MCAT it must satisfy two characteristics: 1) the task must be hearing dependent and 2) failure to perform the task to a specified level must result in decreased safety, efficiency and/or operational effectiveness. For further details about HCTs and MCATs the reader is referred to Chapter 2, sections 2.4 and 2.5. The MCATs carried out by military personnel are currently unknown; in order to develop a measure of AFFD which is based on job specific HCTs for military personnel the MCATs need to be identified. It was decided that at this stage, rather than investigating the MCATs carried out by the entire military population, the focus would be on infantry and combat-support personnel. The skills and tasks carried out by this population are the foundations of the initial training for many other military groups and therefore initially concentrating on this group is considered a sensible starting point.

Study 1 Part A (Section 3.3) uses a focus group approach to gather information about the auditory tasks carried out by infantry personnel; this work was a collaborative effort between the author and Zoë Bevis (PhD Student University of Southampton, see Section 3.3 for further details). Study 1 Part B (Section 3.4) uses a questionnaire to gain a better understanding about which of these auditory tasks can be termed MCATs and which MCATs should be prioritised for representation in a measure of AFFD; this work was carried out solely by the author. The results from Study 1 part A and part B have both been published in Noise and Health Journal (part A: Bevis et al, 2014; part B: Semeraro et al, 2015).

3.2 Research objective 1

Knowledge gap: Measures of AFFD should assess performance on MCATs. There is currently no list of MCATs carried out by infantry and combat support personnel.

Research objective 1: To identify the auditory tasks carried out by infantry and combat support personnel, investigate which of these auditory tasks can be defined as MCATs and to decide which of the MCATs should be prioritised for representation in a measure of AFFD.

3.3 Study 1 part A: exploring auditory tasks

NB. Before reporting Study 1 part A it is worth noting that this piece of research was a collaborative effort between the author and Zoë Bevis (PhD Student, University of Southampton). The author and Zoë Bevis were equally involved in the methodological design, data collection and data inputting (transcribing all the focus groups). The author had limited input with regards to the analysis of the results, taking up an advisory role throughout the analysis process. For this reason a full report of the focus group study is not included in this thesis. The reader is referred to the article by Bevis and Semeraro et al (2014) and Zoë Bevis' thesis (upon completion) for further details about the study. Section 3.3 provides an overview of the motivations, methods, results and conclusions. The emphasis here is specifically on the auditory tasks identified from the focus groups.

The focus group stage of the identification of MCATs aimed to gather a wide range of information about the auditory tasks carried out by infantry personnel and the environment that these tasks are performed in. Before trying to identify a list of MCATs it was necessary to gain a greater understanding of the hearing requirements of infantry personnel. An additional aim of the study was to investigate the underlying attitudes and behaviour of personnel towards noise exposure, hearing loss and hearing protection devices. However the findings on these topics are not reported here, as this information has no direct influence on the AFFD test development (see Bevis et al, 2014 for details).

The focus group method was chosen as the most appropriate method of job analysis as it allows participants to raise relevant issues, discover areas of agreement and disagreement and reflect on past experiences (Pearn & Kandola, 1988). The influence of the researchers is minimised by encouraging the participants to lead the conversation and to discuss topics as a group. A group conversation increases the likelihood of participants raising views that they might not feel comfortable expressing in a one-to-one interview approach. In addition, the method was simple and cost effective, allowing for a large sample size (Pearn & Kandola, 1988; Kitzinger, 1995).

A guideline structure for the focus groups was designed which consisted of seven open-ended questions (see Table 3.1). These questions were developed in consultation with subject-matter experts (SMEs) at the Institute of Naval Medicine, Gosport. The SMEs were a mixture of psychologists who had prior experience of running focus groups with military personnel and defence audiologists who had an understanding of auditory topics within the military. The questions covered topics regarding the auditory tasks performed whilst on tour, the effect of hearing loss on performance of auditory tasks, sources of background noise and hearing protection. All the focus groups began with an introduction about the purpose of the study.

Following this the open-ended questions were used to keep the group discussion relevant to the research aims. The questions were asked in no particular order so as to not disrupt the flow of conversation. Participants were often asked to expand upon ideas they mentioned and the discussion ended when participants had no further information to add to the conversation. All of the discussions were audiotaped and transcribed verbatim.

A total of 16 semi-structure focus group interviews were carried out with eighty British Army personnel, recruited from five infantry regiments across the South of England. Ethical approval for this research was obtained from the University of Southampton (ERGO ref: 5850) and MoD Research Ethical Committee (Ref:359/GEN/12). The mean group size was five, with a range of three to six personnel. All participants had experience of active service and had returned from an operational tour of duty abroad within two months of the study commencing. The sample represented a range of different ranks and infantry occupations. Further participant information can be found in the associated paper (Bevis and Semeraro et al, 2014).

The data was analysed using a typical content analysis method which involved highlighting participant's ideas and opinions in the transcribed data. The analysis aimed to identify qualitative themes which occurred throughout the data. The result of this process was a list of themes that emerged from the ideas and opinions identified. Nvivo 10 (QSR International Pty Ltd, 2012) was used to carry out the data analysis. A second coder was then asked to recode a sample of the data using the original coding descriptions to examine the reliability and objectivity of the coding process. Inter-rated agreement was calculated using the Cohen's Kappa measure and strong agreement (Cohen, 1960; Fleiss et al, 2003) was found between coders ($\kappa = 0.80$).

The analysis resulted in two main themes and seven subthemes (see Table 3.2). The first of the main themes describes the auditory tasks that personnel are expected to perform as part of their operational duties. From within this theme 17 auditory tasks carried out by infantry personnel have been identified (see Table 3.3) which can be divided into the three sub- themes of sound detection, speech communication and sound localisation. The second main theme encompasses factors that personnel believe compromise their performance on auditory tasks. The second theme contains four sub-themes which describe the situations where personnel felt that their hearing ability was reduced or hindered.

Further details regarding all of the themes and subthemes can be found in Bevis and Semeraro et al (2014). The second subtheme 'reasons for reduced performance' is not discussed any further in this report since the focus here is on the auditory tasks identified in the study and what steps need to be taken in order to determine whether these tasks can be deemed MCATs.

The authors commented on which of the tasks mentioned by the participants could be deemed hearing dependent. Any tasks mentioned during the focus groups which began with the detection of a sound, be it speech, weapons firing or footsteps could be seen as a tasks in which hearing is required. Tasks were termed hearing dependent if it was concluded that they could not be carried out using job experience or other sensory modalities alone.

Rather than only focusing on the auditory tasks that were mentioned most frequently the authors decided to report any auditory tasks that were mentioned more than once during the focus groups. This ensured that the list of tasks had not been in anyway 'preselected' by the authors and was instead a complete representation of the ideas and opinions of the participating infantry soldiers.

The auditory tasks identified were discussed abundantly in all the focus groups. The results showed all infantry personnel are expected to be able to carry out these tasks, regardless of their role or rank. The frequency of task performance was, however, shown to be influenced by role. For example, senior personnel and mounted infantry are more likely to communicate via radio than dismounted soldiers or lower ranked personnel. Sound localisation tasks are rarely carried out by those working in engineering roles. Engineers are, however, expected to detect potential vehicle faults from the sound of the engine, a task which dismounted infantry are less likely to carry out. Due to anonymity of the data it is difficult to make clear connections between specific roles and auditory tasks, limiting the generalisability of the data.

From the focus groups a list of the auditory tasks carried out by infantry personnel has been generated (Table 3.3). A further outcome of the qualitative study is an insight into the reasons for reduced performance on auditory tasks which have not be detailed in this report, further information about this aspect of the work can be found in Bevis and Semeraro et al (2014).

It is important to note that one key limitation to the focus group method of job analysis is that the results are entirely reliant on the topics that personnel deemed important to mention. As a result, it may be that there are additional auditory tasks that personnel did not mention. One topic worth mentioning, that did not arise in the focus groups, is the importance of 'acoustic stealth'. Stealth is defined as the "cautious and surreptitious action or movement" (Oxford English Dictionary, 2015). It may be the case hearing impairment affects personnel's ability to maintain acoustic stealth in close combat situations but this was not mentioned in response to questions two and three in Table 3.1. The enemy may detect your location by sounds you make, such as footsteps, rustling of clothing or heavy breathing. The importance of this skill in an operational environment, and its impact of AFFD, is unknown. This is be further investigated by Matthew Blyth as part of his PhD. Given the variety of ranks, roles and experience for the personnel who

participated, the author is satisfied that the list of acoustic tasks obtained in Study 1 part A is, on the whole, representative of the tasks carried out by infantry and combat-support personnel. However, carrying out an observational job analysis may reveal additional tasks that were not mentioned in the focus groups and is recommended for future work identifying MCATs for other military roles and HCTs for other occupations.

At this stage it is not possible to report which of the auditory tasks identified in Study 1 part A satisfy the characteristics of a MCAT. It was judged by the author that all 17 auditory tasks identified in the focus groups cannot be carried out using job experience or other sensory modalities alone and are therefore hearing dependent, the first characteristic of an MCAT. The term hearing dependent simply implies that to carry out these tasks the audition of a sound is required. This does not mean to say that normal hearing personnel are in fact able to successfully carry out these tasks using hearing alone; this will need to be further investigated for each task. However, at this stage it is not possible to report on the criticality of the tasks, the second characteristic of a MCAT. The second stage of this work aims to address this (Study 1 part B, reported in Section 3.4).

Table 3.1 *List of focus group questions*

1	Can you describe the types of noise you were exposed to on tour?
2	Describe any situations whilst performing your job in which you thinking having good hearing is critical.
3	Can you recall any time when you have been unable to hear clearly when performing your role?
4	Can you recall a situation when you were unable to make yourself heard?
5	Can you describe the impact, if any, that your hearing protection has on your ability to hear whilst on tour?
6	How do you communicate important signals with each other?
7	Can you describe any situations where determining the location of a sound source was important?

Table 3.2 *Focus groups: themes and sub-themes*

Themes	**Sub-themes**
1. Auditory Tasks	1.1 Speech Communication
	1.2 Sound Localisation
	1.3 Sound Detection
2. Reasons for reduced performance	2.1 Background noise
	2.2 Hearing protection devices
	2.3 Stress
	2.4 Attention difficulties

Table 3.3 *Focus group: theme one- infantry auditory tasks*

Sub-theme	**Tasks**
Sub-theme 1.1 **Speech communication**	T1 Hearing commands in a casualty situations T2 Hearing grid references T3 Hearing directions on patrol T4 Hearing directions in a vehicle T5 Hearing fire control orders T6 Hearing stop commands T7 Hearing the briefing before a foot patrol T8 Communicating through an interpreter
Sub-theme 1.2 **Sound localisation**	T9 Locating a small arms firing point T10 Locating an artillery firing point T11 Locating the moving sound source of a motorbike T12 Locating the moving sound source of footsteps T13 Locating enemy movement in maize fields T14 Locating a talker
Sub-theme 1.3 **Sound detection**	T15 Identifying the type of weapon systems being fired T16 Determining talker identity T17 Detecting a malfunction in an item of machinery

3.4 Study 1 part B: identification of MCATs

3.4.1 Introduction

Study 1 part A (Section 3.3) reports a list of 17 auditory tasks carried out by infantry and combat-support personnel whilst on operational duties. These are listed in Table 3.3. The tasks have been split into three categories, speech communication, sound detection and sound localisation. Although the data from Study 1 part A provides information about the complex auditory environment British infantry personnel are working in, it is not possible to use this qualitative data to produce a list of MCATs. It has been acknowledged that measures of AFFD should measure performance on job specific HCTs (see sections 2.4 and 2.5). Therefore, in order to develop a measure of AFFD for military personnel it is necessary to identify which of the 17 auditory tasks can be considered MCATs. Further information is needed about each task in order to determine which of the auditory tasks satisfy the two characteristics of a MCAT.

The first characteristic of an MCAT is hearing dependency. It was judged by the primary author of the present study that all 17 auditory tasks identified in Study 1 part A require the audition of a sound and cannot be carried out using job experience or other sensory modalities alone; the tasks have therefore been deemed hearing dependent. The second characteristic of an MCAT is that failure to perform the task to a specified level will result in decreased safety, efficiency and/or operational effectiveness. To determine whether an auditory task meets this criterion, knowledge of the consequences of poor performance is needed.

Following the identification of MCATs, a measure of auditory fitness can be created or adapted to represent all or a selection of these tasks. Representing a task does not imply that an exact replication of the MCAT should be included as part of the AFFD test battery, but infers that the auditory skills personnel require to carry out the task should be assessed. Measures of AFFD should accurately assess performance on HCTs and should be generally applicable to the majority of employees within a given occupation. In a military context, a suitable AFFD test battery needs to include performance measures that are appropriate for the majority of ranks and roles. A compromise is needed to accurately measure auditory fitness on specific tasks without creating a test battery that is valid only for a small proportion of personnel. Documenting who performs the MCATs, and how frequently, highlights tasks that are seldom carried out or those performed by small numbers of personnel. These tasks do not need to be prioritised for representation in a measure of AFFD.

To summarise, three pieces of information about each auditory task are required in order to determine which of the tasks are mission-critical and which should be represented by a measure of AFFD: (1) the consequences of poor performance on the task; (2) which ranks and roles perform the task; (3) how frequently the task is performed. Study 1 part B aims to gather this information for each of the 17 auditory tasks.

For this study, one-to-one interviews, focus groups and questionnaires were considered as data collection techniques. One-to-one interviews or focus groups with infantry and combat-support personnel (as used in Study 1 part A) can be used to explore participants' thoughts and opinions in detail, and to discover areas of agreement and disagreement (Kitzinger, 1995). However, these methods typically produce unstructured data, which (for the purpose of this study) would need to be organised, coded and quantified, and are also prone to researcher bias. Questionnaires, conversely, do not allow for such detailed exploration, but can provide more readily quantifiable data.

A questionnaire approach was adopted by Brown and Fallowfield (2012) in their work on developing a strength-based Royal Navy Fitness Test. They first created a list of strength-based tasks performed on board Royal Navy ships through consultation with subject-matter experts and then used a questionnaire approach, with Likert type scales to collect information about the strength demands, importance and frequency of each task. This produced quantifiable data that were used to identify the most critically-demanding generic tasks performed on board Royal Navy ships. This style of questionnaire was adopted by the authors in the present study.

3.4.2 Aims

The present study aimed to identify which of the auditory tasks carried out by infantry and combat-support personnel can be defined as MCATs and which MCATs should be represented by a measure of AFFD.

3.4.3 Methods

The list of tasks included in the questionnaire was taken directly from the focus group results listed in Table 3.3). For each of the 17 auditory tasks participants were required to give Likert scale ratings concerning (1) the significance of the consequences of poor performance; (2) whether the task is carried out by all, some or no infantry personnel; (3) how frequently the task is performed during a training exercise or when serving on a tour of duty. The questionnaire was developed in consultation with subject-matter experts at the Institute of Naval Medicine, Gosport

(the same individuals who provided advice for the development of the focus group guideline structure, Section 3.3).

The response options for the 'consequences of poor performance' and 'frequency of task' questions were based on the scales used for assessing the risk of events on generic risk assessment documents (Health and Safety Executive, 2014). To determine the consequence of poor performance, the consequence scale used in the University of Southampton Risk Estimation Matrix (University of Southampton, 2013) was used. For the question relating to frequency of task performance, the corresponding scale in the Royal Navy physical strength questionnaire (Brown & Fallowfield, 2012) was used. Response options for who performs each task were limited to 'all', 'some' or 'no infantry personnel'. If the participant selected 'some' they were asked to indicate which roles carried out that particular auditory task. The options for each question are shown in Table 3.4. Due to the range of auditory tasks it was not possible to provide descriptions for response answers (for example what is meant by 'minor consequence') that would be applicable to all tasks. Participants were therefore not given guidance on how to interpret the answers to each question.

Participants were recruited from four regiments across the South of England. The questionnaire and a covering letter outlining the study were sent via email to eleven senior personnel who had been involved in Study 1 part A. A total of seven responses were received, and four responses were positive (see Appendix O for details of recruitment challenges in the military population). Four regiments completed the questionnaire, resulting in a total of 87 questionnaire responses (Regiment One n=34; Regiment Two n=16; Regiment Three n=23; Regiment Four n=14). All participants had experience of an infantry or combat-support role, either during training exercises or during an operational tour of duty and represented a wide range of ranks and roles. Participant details are given in Table 3.5.

The senior personnel were tasked with organising voluntary groups of infantry personnel, regardless of rank or role, resulting in an opportunistic sampling approach. The researcher requested that the senior personnel distribute the participant information sheet and consent forms 24 hours before data collection, giving participants opportunity to withdraw from the study if they wished. The questionnaires for one regiment were distributed and collected by the author. For the remaining three regiments, senior personnel distributed and collected the questionnaires, and forwarded them to the first author.

Data collection took place at the participants' normal place of work. Prior to giving consent and completing the questionnaire it was reiterated that participation was voluntary and that personnel could withdraw at any time without giving a reason. Consent forms were not attached

to the questionnaires, ensuring that all responses were anonymous. In total, 87 questionnaires were completed and 79 were used for analysis. The questionnaire took between ten to twenty minutes to complete. Reasons for questionnaire exclusion were: incomplete questionnaire (n=4); incorrect use of scale, for example answering '4' when the options were numbered 1-3 (n=3); giving the same answer for every question indicative of the instructions not being followed (n=1).

Within Regiment Four a small group of participants (n=8) were asked to complete the questionnaire a second time five days later to collect data on the repeatability of the questionnaire. The participants were each given a number to write on both of their questionnaires, making it possible to link the data whilst ensuring anonymity. This group was selected using opportunistic sampling and only participants that would be available to fill in the questionnaire on both dates were selected.

Ethical approval was obtained for this study from the University of Southampton (ERGO ref: 6686) and MoD Research Ethical Committee (Ref:359/GEN/12). All data collected were anonymous and treated confidentially.

The results from all participants were pooled and are reported as median values or as a proportion of all responses. The data gathered from the Likert-type scale data are ordinal and therefore the median is the most appropriate measure of central tendency (Boone & Boone, 2012).

Table 3.4 *Questionnaire survey guide*

CONSEQUENCES of poor performance	WHO performs this task?	FREQUENCY of task
In your opinion how significant are the consequences of poor performance on this task?	In your opinion, during a training exercise or when serving on a tour of duty is this task carried out by all infantry personnel, some infantry personnel or no infantry personnel?	In your opinion, how frequently is this task performed during a training exercise or when serving on a tour of duty?
1 = No Consequence 2 = Minor 3 = Moderate 4 = Major 5 = Critical	1 = No infantry personnel 2 = Some infantry personnel (indicate which roles) 3 = All infantry personnel	1 = Seldom or yearly 2 = Occasionally or monthly 3 = Regularly or weekly 4 = Frequently or daily 5 = Continuously or several times per day

Table 3.5 *Study 1 part B participant information*

Characteristics	Number of participants n (%)
Gender	
Male	78 (99)
Female	1 (1)
Rank	
Private	36 (45)
Lance Corporal	10 (13)
Corporal	12 (15)
Sergeant	10 (13)
Warrant Officer	1 (1)
Lieutenant	3 (4)
Captain	5 (6)
Major	2 (3)
Number of tours of duty	
0	12
1	25
2	16
3	8
≥4	18
Tour locations	
Afghanistan	94
Northern Ireland	11
Iraq	29
Macedonia	1
Bosnia	14
Cyprus	2
Kosovo	7
Falklands	1
Not stated	23
Time serving in the Armed Forces (completed years)	
Mean (min/max)	8 (1/27)

3.4.4 Results: consequences of poor performance

Table 3.6 shows the proportion of responses for each consequence rating with the median rating shaded black. For all the speech communication tasks (T1-T8, see Table 3.3 for list of tasks) the majority of the responses (>68%) indicated that poor performance would result in a critical or major consequence, with very few participants rating the tasks as having less than a moderate consequence. For the sound localisation tasks only T9, T10 and T12 received the majority of responses (>65%) for the critical or major consequence categories. For the sound detection tasks only T15 was rated as having critical or major consequence by the majority of participants (53%). For all the tasks the median consequence score is above 3 (moderate consequence); this may be a result of response bias and is considered further in the Discussion (Section 3.4.9).

Table 3.6 *Percentage of responses for each task for each consequence rating. The shaded black section shows the median rating. Each column totals 100%.*

Consequences of poor performance	**Task *(proportion of responses, %)***																
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
No consequence *(1)*	1	0	3	0	1	0	2	3	3	2	3	1	4	4	3	2	5
Minor consequence *(2)*	1	2	4	4	2	1	0	0	0	6	25	6	20	14	5	25	15
Moderate consequence *(3)*	7	13	25	24	10	7	18	24	11	27	**33**	18	**40**	**38**	39	**37**	**33**
Major consequence *(4)*	14	**37**	**25**	**35**	22	27	**34**	**43**	25	**22**	23	**29**	18	23	**34**	27	28
Critical consequence *(5)*	**77**	48	43	37	**65**	**65**	46	30	**61**	43	16	46	18	21	19	9	19

Table 3.6 shows that there was a large variation in the responses to the question about negative consequences from poor performance. For all of the tasks, a few participants (≤5%) responded with 'no consequence'. This may be due to participants not reading the question or scale correctly or misinterpretation of the question, as opposed to a genuine belief that poor performance on a task would have no negative consequence. For example, it seems unlikely that personnel would suggest there would be no consequences if directions on a foot patrol were not accurately heard, yet 2.5% of participants responded with this answer. This is covered further in the Discussion (Section 3.4.9).

Despite a few personnel responding 'no consequence', for all the tasks there is a general consensus amongst participants that there is some consequence of poor performance, with the majority of responses (≥95%) of all the tasks falling within the range of minor to critical consequence. Therefore, it can be concluded that all 17 auditory tasks can be considered 'mission-critical'.

3.4.5 Results: roles that carry out each task

The majority of participants responded that all infantry personnel are expected to carry out every task, as shown in Table 3.7. The median result for each task, apart from one, was a score of 3, indicating that there was a general agreement between participating personnel that these tasks are carried out by all personnel. Only T8 (communicating accurately through an interpreter) had a lower median score of 2 (carried out by some infantry personnel); for this reason it was not deemed necessary to represent T8 in a measure of AFFD for infantry and combat-support personnel.

Table 3.7 *Percentage of responses for each task for the question 'who performs this task'. The shaded black section shows the median rating. Each column totals 100%.*

Who performs the task	**Task *(proportion of responses, %)***																
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
No infantry personnel *(1)*	5	4	4	3	3	3	4	6	3	4	3	3	6	3	3	10	13
Some infantry personnel *(2)*	14	33	9	39	19	10	10	46	5	11	6	6	3	6	10	19	10
All infantry personnel *(3)*	**81**	**63**	**87**	**58**	**78**	**87**	**86**	48	**92**	**85**	**91**	**91**	**91**	**91**	**87**	**71**	**77**

If the participants answered 'some infantry personnel' for any task they were asked to indicate which roles carried out the task. Tasks T2, T4 and T8 gained the highest number of responses for only some infantry personnel carrying out the tasks. The roles that were said to carry out each of these tasks are listed in Table 3.8; it is apparent that a number of different roles perform each task.

Table 3.8 *Infantry personnel reported to perform tasks 2, 4 and 8; the three tasks with the highest number of responses for 'some infantry personnel'. Number in brackets indicates the number of responses given for each role*

Task 2 Accurately hearing grid references *(33% of responses for 'some infantry personnel')*	**Task 4 Accurately hearing directions in a vehicle** *(39% of responses for 'some infantry personnel')*	**Task 8 Communicating accurately through an interpreter** *(46% of responses for 'some infantry personnel')*
Those in command (23) Those who communicate over radio (2) Point man (1) Signaller (1)	Those who drive vehicles (14) Those in command (12) Gunners (5) Those in vehicle mounted regiments (4) Fire Support Groups (1) Dismounted Commander (1)	Those in command (29) Interpreters (4) Ground personnel only (1)

Whilst the majority of participants answered that all tasks are carried out by some or all infantry personnel, there were a small number (≤13% for any given task) that responded with 'no infantry personnel'. It is surprising that these participants responded in this manner given that the majority of tasks seem pivotal to the infantry role, for example 'accurately hearing directions on a foot patrol'. This raises concern that some participants were not clear about the meaning of the question; this is addressed in the Discussion (Section 3.4.9).

3.4.6 Results: frequency of task performance

Table 3.9 shows the proportion of responses for each frequency rating with the median values shaded black. None of the tasks had a median frequency rating of 5 (continuously or several times per day). Five tasks (T2, T3, T4, T5 and T7) had a median frequency rating of 4 (frequently or daily); these five tasks were all speech communication tasks. For all tasks apart from two (T10 and T17), the majority of responses (>50%) indicate that the tasks are carried out 'regularly or weekly' or more frequently.

Surprisingly, 5% of participants responded that accurately hearing grid references was a task carried out seldom or yearly when the majority responded that this task was continuously or several times per day. This result may mean that some participants were answering the questionnaire based upon their individual role within the infantry as opposed to the infantry as a whole.

Table 3.9 *Percentage of responses for each frequency rating. The shaded black section shows the median rating. Each column totals 100%.*

Frequency of task performance	**Task (proportion of responses, %)**																
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
Seldom or yearly *(1)*	7	5	6	4	3	4	6	17	6	32	9	18	14	10	9	20	21
Occasionally or monthly *(2)*	35	15	14	11	24	26	20	15	24	**27**	32	**32**	30	23	20	29	**35**
Regularly or weekly *(3)*	**38**	17	9	24	23	**30**	17	**21**	**24**	23	**30**	29	**27**	**27**	**39**	**23**	24
Frequently or daily *(4)*	14	**24**	**26**	**26**	**26**	17	**36**	24	29	11	20	17	21	17	26	12	14
Continuously or several times per day *(5)*	6	39	45	35	24	23	21	23	17	7	9	4	8	23	6	16	6

3.4.7 Results: identifying tasks to be represented in a measure of AFFD

To determine which tasks are most important for an AFFD measure, the tasks were arranged in a consequence/frequency matrix. They have been positioned according to their median consequence and frequency score (Figure 3.1).

There is no definitive way to combine these two pieces of data since no previous research has been conducted in this area. Grey-scale classification has been used to represent the importance of the task being represented in an AFFD assessment. The tasks in the black area are those that should be prioritised. The tasks in the grey and white areas are performed less frequently and/or have lesser consequences, causing them to be of lower priority.

Figures 3.2 and 3.3 show the percentage of participants whose responses agree with the matrix (Figure 3.1). The error bars show the 95% confidence intervals of the participant's responses for the two consequence rating groups (see key on Figure 3.2) and the two frequency rating groups (see key on Figure 3.3). Figure 3.2 shows that the majority of personnel (over 70% across all tasks) stated that there would be at least a moderate consequence for all tasks if performed poorly. The large gap between the two consequence groupings suggests that participants were in agreement about the consequence of poor performance on each task. Figure 3.3 shows that for all the tasks that fall in the black area of the matrix (Figure 3.1) the majority of personnel (over 55% across all tasks) stated that these tasks are carried out at least regularly or weekly.

		Consequences of poor performance				
Frequency of task performance		**No consequence**	**Minor**	**Moderate**	**Major**	**Critical**
	Seldom or yearly					
	Occasionally or monthly			**T17$**	**T10~ T12~**	
	Regularly or weekly			**T11~ T13~ T14~ T16$**	**T15$**	**T1* T6* T9~**
	Frequently or daily				**T2* T3* T4* T7***	**T5***
	Continuously or several times per day					

Figure 3.1 *Consequence/frequency matrix (Key *speech communication, ~sound localisation, $sound detection), numbers relate to tasks in Table 3.3*

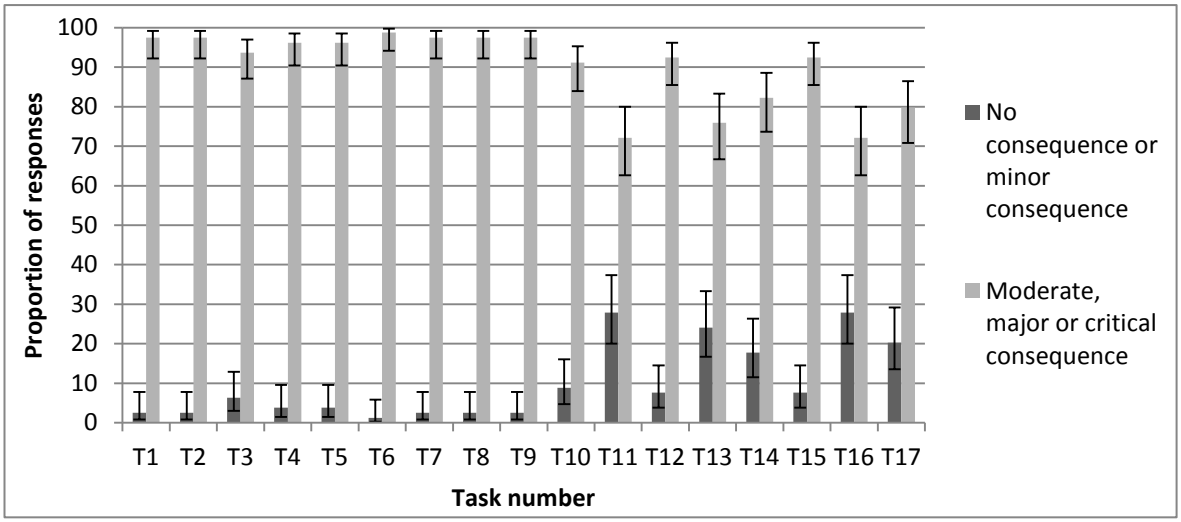


Figure 3.2 *Percentage of responses for no/minor consequence if the task is performed poorly and moderate/major/critical consequence if the task is performed poorly. Error bars show 95% confidence intervals.*

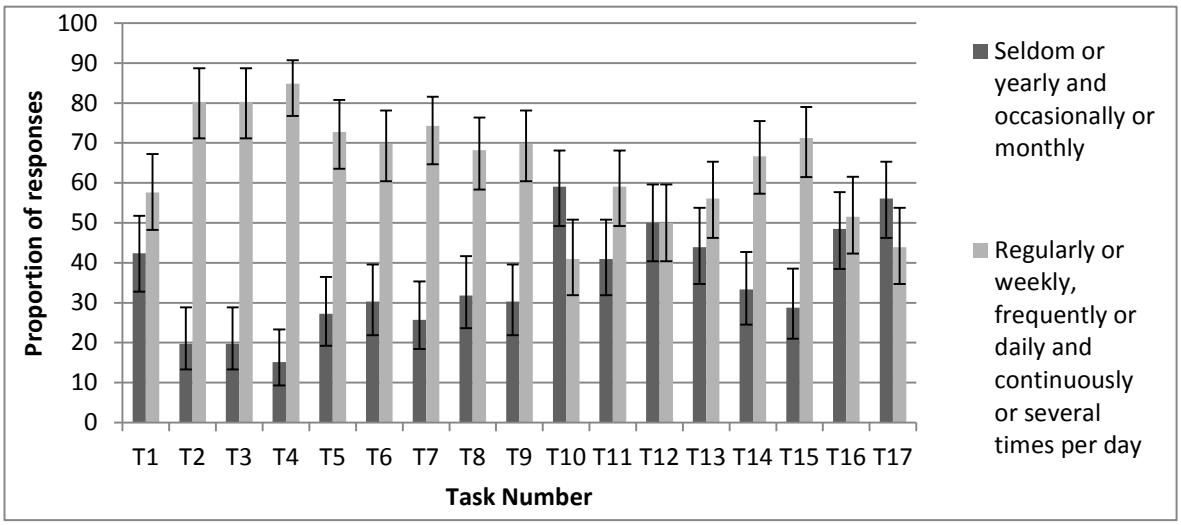


Figure 3.3 *Percentage of responses for tasks performed seldom or yearly/occasionally or monthly and those performed regularly or weekly/frequently or daily/continuously or several times per day. Error bars show 95% confidence intervals.*

Using this matrix approach assumes that the consequence and frequency ratings have equal weighting (a task carried out regularly or weekly with major consequence falls in the same category as a task carried out frequently or daily with moderate consequence). There is one area of the matrix where this assumption may be problematic. Firstly, tasks carried out seldom or yearly with critical consequence should arguably fall into the black area; personnel should be able to carry out any task that has critical consequence, even if they are rarely required to perform it. Although this is an important consideration when using a risk matrix, during the present study no tasks fell within this area.

The decision to place the cut off point for inclusion in a measure of AFFD between the grey and black area is arbitrary. This point was chosen in order to include the tasks with

moderate/major/critical consequence if the task is performed poorly and tasks performed regularly or weekly/frequently or daily/continuously or several times per day. By using the matrix in Figure 3.1 it is possible to generate a list of MCATs that should be prioritised for representation in a measure of AFFD for infantry personnel and those in combat-support roles; these are listed in Table 3.10.

Table 3.10 *List of MCATs to be prioritised for representation in a measure of AFFD for infantry personnel*

Speech communication	T1: Accurately hearing commands in a casualty situation
	T2: Accurately hearing grid references
	T3: Accurately hearing directions on patrol
	T4: Accurately hearing directions in a vehicle
	T5: Accurately hearing fire control orders
	T6: Accurately hearing 'stop' commands
	T7: Accurately hearing the briefing before a foot patrol
Sound localisation	T9: Locating a small arms firing point
Sound detection	T15: Identifying the type of weapon systems being fired

3.4.8 Results: test-retest reliability

In order to estimate the repeat-reliability of the questionnaire a sample of the participants (n=8) completed the questionnaire twice. The participants were all from Regiment Four. For each participant the responses were summated across each of the questions (consequences, who and frequency). This resulted in six values for each participant representing their responses to each question, on the two questionnaires. Since the data were measured on a Likert type scale, the Spearman's Rank-Order Correlation was selected to measure the strength of association between the repeated questionnaires (Jamieson, 2004). There was a positive correlation between responses on the two completed questionnaires, which was statistically significant ($r_s(22) = .803$, $p = <.001$). The absolute difference between responses on the questionnaire repeats was calculated. Across the eight participants, the average absolute difference in ratings between questionnaire repeats was a change of less than one rating for all the questions (0.6 for the consequences of poor performance, 0.3 for who performs each task and 0.7 for the frequency of task performance).

3.4.9 Discussion

The primary aim of this study was to identify MCATs for infantry and combat-support personnel. The secondary aim was to determine which MCATs should ideally be represented by a test of AFFD. A list of 17 auditory tasks carried out by infantry personnel and combat-support roles were taken from Study 1 part A (Section 3.3) and were further investigated using a questionnaire. Each

task was rated with regards to the consequences of poor performance, who carries out the task and the frequency of task performance.

All 17 of the tasks from Study 1 part A can be considered MCATs; they are all hearing dependent and poor performance could result in a significant negative consequence. Distinguishing between 'significant consequence of poor performance' and 'non-significant consequence of poor performance' is subjective and for this reason a cut-off was not used. However, from Table 3.6 it can be seen that none of the tasks were rated by the majority of participants as having no consequences to poor performance. It is assumed that any consequence could have a negative impact on the safety, efficiency and/or effectiveness of the task in question. Therefore, all 17 tasks can be classed as MCATs.

It has been established that a measure of AFFD should be based on job specific tasks (Tufts et al, 2009; Laroche et al, 2003) and in a military specific context this means basing an AFFD test on MCATs. All of the MCATs identified in the present study are eligible for inclusion when designing a measure of AFFD for infantry personnel. It is not proposed that a measure of AFFD includes exact replicas of the tasks that are performed in the MCATs identified. For example, performance on the task 'accurately hearing the briefing before a foot patrol' does not necessarily have to be assessed by replicating the task itself. The idea is to assess the types of auditory skills needed to perform well in that environment, such as hearing speech in the presence of background noise or being able to localise a sound source.

In order for a task to be prioritised for representation by a measure of AFFD that is applicable for the majority of roles and responsibilities, it is first important to establish that the task is carried out by the majority of infantry personnel. The data from Question Two, 'Who performs this task?' were used to exclude tasks that were only carried out by specific roles. Only one task (T8, communicating accurately through an interpreter) was found to be carried out by some, as opposed to all, infantry personnel and was therefore excluded from further consideration.

The final stage of the analysis involved identifying the tasks that are performed frequently and have significant consequences when performed poorly. Tasks that are performed infrequently and/or have minor consequences to poor performance were considered as having low priority for representation in a measure of AFFD. The results in Figure 3.1 show that only one of the tasks (T17, detecting a malfunction of an item of machinery) falls within the white area, indicating low frequency of performance and minor consequences of poor performance. By only incorporating tasks that fall into the black area of the frequency/consequence matrix (Figure 3.1) the AFFD measure will represent the tasks that are performed most frequently and have the most significant effect on the safety and effectiveness of a mission.

Tables 3.6 and 3.9 display the variation in participant responses for the consequence and frequency data. They show that there is a large variation in the answers given for the questions about frequency and consequence. However, for the all tasks that fall within the black area of Figure 3.1 the majority of participants responded that the tasks have moderate to severe consequence (ranging from 92% of responses for T15 to 99% of responses for T6) and are carried out regularly or weekly to continuously or several times per day (ranging from 58% of responses for T1 to 85% of responses for T4).

For all three questions there were some unexpected responses from a small number of participants which may be due to misinterpretation of the questions. It seems unlikely that a participant would state that 'no infantry personnel are required to accurately hear directions on a foot patrol' since this is an integral part of being an infantry soldier, yet 3.8% of participants gave this response. Similarly, unexpected responses were given for the consequence and frequency questions. For example, 2.5% of participants answered that there are no consequences when a small arms firing point is not located (or located incorrectly; T9) and 4.5% responded that accurately hearing grid references (T2) is a task carried out seldom or yearly. These unexpected responses call into question how participants interpreted the questionnaire. It could be that some individuals were answering based solely on their role within the Armed Forces, rather than considering the general role of an infantry soldier. For an individual who has only been serving for a short time period or has not yet been on a tour of duty it is possible that they do not have sufficient experience to call upon. They may not have experienced the consequences of poor performance on a task, or may not perform certain tasks as part of their role, causing these individuals to answer that there is no consequence if the task is poorly performed or that the task is rarely performed. The questionnaire was intended to yield information about the infantry workforce in its entirety, not about individual roles or experiences.

It is not known if the order of the questions had any influence on the data obtained. It is possible that participants' views on what each scale item meant to them evolved as they filled out the questionnaire, causing their opinions to change when answering questions towards the end of the questionnaire in comparison to the beginning (Bowling, 2005).

Individual decision criteria would have varied between participants, particularly when answering the consequence question. It is possible that one participant considered the injury of a colleague as a 'critical' consequence, whereas another participant may not consider the situation critical until there is loss of life. This may account for the variation in participant responses for this question.

Finally, the phrasing of the question regarding the significance of consequences of poor performance could be interpreted to imply that the task is not inconsequential. Participants could, therefore have been discouraged from selecting option one, 'no consequence' and possibly even option two, 'minor consequence', resulting in a response bias.

The final list of MCATs (Table 3.10) includes all three types of auditory task: speech communication, sound localisation and sound detection. These findings are consistent with those of Tufts et al (2009) who recognised that an AFFD test battery should include measures of functional hearing ability including speech understanding, sound localisation, sound detection and recognition. The end result of this questionnaire is an objective, evidence-based list, characterising the most important tasks to be represented by a measure of AFFD for infantry personnel. This is the first step towards developing a measure of AFFD based on the jobs carried out by infantry and combat-support personnel. There are auditory tests that measure performance on these aspects of hearing ability, such as speech threshold testing and sound source identification tasks. However, there are currently no auditory tests used by the UK military that have been validated to measure AFFD.

As the focus groups and questionnaire were carried out using participants recruited from the infantry (Army and Royal Marines) and combat-support roles, the results from this study cannot be generalised to the wider Army, Royal Air Force or Royal Navy. Whilst the specific findings cannot be applied to other populations, this methodological approach can be applied to other occupations, both within the MoD and other workplace environments, where a measure of AFFD based on job-specific HCTs is required. Collecting information about auditory tasks directly from employees ensures that the results are a true reflection of the occupation in question.

Using the results from Study 1 parts A and B it can be hypothesised that other military cohorts also carry out auditory tasks requiring speech communication, sound localisation, and sound detection auditory skills. These professions also use PTA as their primary auditory screening method and it is therefore suggested that a similar study is carried out for these cohorts. This would determine which specific auditory skills should be tested as part of their AFFD measurements.

3.4.10 Conclusion

This study has produced a list of 17 MCATs carried out by infantry and combat-support personnel in the British Army. Nine of these MCATs are performed by the majority of ranks and roles either weekly or daily and have either major or critical consequence if performed poorly. These nine MCATs should be prioritised for representation by a measure of AFFD for infantry and combat-

support personnel to ensure they have the necessary auditory skills for safe and effective deployment on operational duties.

3.5 Chapter 3 summary

Study 1 (parts A and B) has explored the auditory tasks carried out by infantry and combat support personnel and a list of 17 MCATs performed by these personnel has been produced. A measure of AFFD should be based on the auditory tasks required to carry out the job in question. Nine of the 17 MCATs have been identified as being performed by the majority of ranks and roles, either weekly or daily and have either major or critical consequence if performed poorly. The auditory skills required to perform these nine MCATs (see Table 3.10) should be prioritised for representation by a measure of AFFD. These auditory skills can be categorised into three broad categories: speech communication, sound localisation and sound detection.

The current measure of AFFD used within the military is PTA but there is some concern about whether this measurement tool is a suitable measure of AFFD. The relationship between performance on the speech communication MCATs (SC-MCATs) and an individual's pure-tone thresholds is currently unknown. At this stage it is therefore not possible to use pure-tone thresholds to predict whether someone is fit for duty. In addition, PTA solely measures the audibility aspect of hearing impairment and none of the additional processing deficits associated with sensorineural hearing loss (Plomp, 1978). It is therefore not known whether PTA is able to accurately predict performance on the MCATs or whether additional measurement tools, such as SIN tests or source identifications tests, should be introduced as new measures of AFFD; this is further explored in Chapter 4.

Before considering the best methods for measuring performance on these MCATs it needs to be established whether normal hearing personnel are in fact able to carry out these tasks. For the SC-MCATs it can safely be assumed that personnel are able to do these tasks using their hearing. However, for the sound localisation and sound detection MCATs there is currently sparse literature assessing whether personnel are in fact able to carry out these tasks. The remainder of this thesis is focused on the SC-MCATs and exploring whether measures of AFFD are able to predict performance on these tasks. A separate PhD project, by Zoë Bevis is focused on the sound localisation MCAT and whether this auditory skill should be assessed when measuring AFFD.

Chapter 4: Is pure-tone audiometry (PTA) fit for duty?

4.1 Introduction

The principal aim of Chapter 4 is to argue why PTA may not be a suitable tool for assessing AFFD, specifically focusing on the speech communication subset of the MCATs (SC-MCATs). Within Chapter 3, a list of MCATs carried out by infantry and combat-support personnel were identified. It has been established that measures of AFFD should be based on measuring whether employees are capable of carrying out job-specific hearing critical tasks (HCTs) safely and effectively (Tufts et al, 2009). It therefore follows that the measure of AFFD used within the Armed Forces should be able to accurately predict performance on the MCATs identified in Chapter 3. This is known as predictive validity, a measure of how well one test can predict a set of abilities on another (see Appendix A for a glossary of types of validity).

As a gold standard, in order to assess the predictive validity of PTA as a measure of AFFD, the association between performance carrying out the SC-MCATs and audiometric thresholds should be investigated. However, if predictive validity can be explored using existing evidence, developing a method for assessing performance on the SC-MCATs would be time-consuming and potentially unnecessary. Chapter 4 outlines the existing evidence relevant to whether PTA is likely to be able to predict performance on the SC-MCATs. This is investigated by addressing two topics. Firstly, PTA is a measure of hearing acuity, not overall hearing ability (Section 4.2). Secondly, there is evidence to suggest that PTA is not very good at predicting performance when listening to complex signals, such as speech (sections 4.3 and 4.4). These two arguments are outlined further below.

PTA is a measure of tone detection in quiet, measuring ‘audibility’. Audibility refers to whether an individual is capable of detecting a sound, i.e. whether it is presented above their hearing threshold. When listening to the SC-MCATs personnel are required not only to detect the speech signal, but also to make sense of what is being heard. This can be referred to as measuring their overall ‘hearing ability’. Auditory ability is an abstract concept (Kidd et al, 2007), in the same way that ‘general intelligence’ and ‘physical fitness’ are not easily quantifiable. It the context of this thesis it is used to refer to an individual’s behavioural hearing, i.e. how well they perform, in general, when carrying out a variety of listening tasks. In comparison to the pure-tones presented during PTA, the speech signals which make up the SC-MCATs are complex, fluctuating in level and frequency over time. The auditory abilities required to listen to these more complex signals are different to those required when simply detecting a tone in quiet. It is understood that

psychoacoustic and non-psychoacoustic abilities that are not assessed by PTA, are required when carrying out speech intelligibility tasks. These factors are explored in Section 4.2. This section aims to show that PTA cannot explain individual differences on speech intelligibility tests, since the psychoacoustic and non-psychoacoustic abilities utilised for detecting a tone-in-quiet are different to those required for SIN tasks. This influences whether PTA should be considered to be an appropriate tool for accurately predicting individual differences in performance when carrying out the SC-MCATs.

The second argument made in Chapter 4 is that there is evidence to suggest that PTA is a poor predictor of performance in complicated listening environments, such as the SC-MCATs (sections 4.3 and 4.4). A review of the literature reporting the association between audiometric thresholds and performance on speech intelligibility tasks is provided in Section 4.4. An assumption is made that the listening skills involved when carrying out speech intelligibility tests are similar to those required for the SC-MCATs (outlined in more detail in Section 4.2). If PTA is not able to accurately predict performance on speech intelligibility tasks then it is assumed that it is unlikely to be an accurate predictor of performance on the SC-MCATs.

Chapter 4 aims to outline why investigating a speech intelligibility task as an alternative tool for assessing performance on the SC-MCATs, is a sensible next step when considering the development of a new tool for predicting AFFD. The key arguments for each section of Chapter 4 are outlined at the start of each section to help guide the reader.

4.2 PTA: A measure of hearing acuity not hearing ability

Section 4.2 key argument
In principle, PTA cannot explain individual differences in performance on speech intelligibility tests.

This section aims to explore the theoretical argument why PTA, a measure of hearing acuity, may not be an accurate predictor of performance in complex listening environments, such as the SC-MCATs. It is obvious that audibility has a crucial role to play in the perception of speech; if the speech signal cannot be heard then information is lost and speech intelligibility is compromised. For hearing impaired individuals (where impairment relates to a reduction in audiometric thresholds) aspects of the speech-spectrum are inaudible, resulting in a reduction in speech intelligibility. However, it is also known that hearing impaired individuals display poorer speech intelligibility ability than normal hearing listeners, even when the level is raised to be well within their audible range (Moore, 1998). Two reasons are given below to suggest that factors other than audibility contribute to the speech intelligibility difficulty experienced by hearing impaired

listeners. It is important to note that the type of hearing impairment being referred to here is of cochlear origin, not a conductive loss.

1. ***The Speech Intelligibility Index (SII) is unable to accurately predict speech intelligibility ability of hearing impaired individuals***

   The SII is a method of quantifying the proportion of a speech signal that is audible and, by weighting the relative importance of different frequencies for speech intelligibility (based on the long term average speech-spectrum), calculating the overall intelligibility of a speech signal (Ma et al, 2009). If all of a speech-spectrum is audible the signal has an SII value of 1.0 (Moore, 1998). The SII is based on the assumption that audible speech is intelligible speech; if this assumption is true then the SII would be capable of accurately predicting speech intelligibility for hearing impaired listeners by using audiometric thresholds to calculate the proportion of the signal that is suprathreshold. Several studies have shown that the SII is able to predict intelligibility fairly accurately for those with normal hearing or a mild hearing loss. However, for more moderate to severe losses the speech intelligibility is worse than that predicted by the SII (Moore, 1998; Moore, 2003). This indicates that for hearing impaired listeners (with a hearing loss of cochlear origin), audibility is not the only factor involved in speech intelligibility; other suprathreshold discrimination abilities are also required, covered in Section 4.2.1.

2. ***Hearing impaired listeners display reduced speech intelligibility ability in comparison to normal hearing listeners even when speech is presented at suprathreshold levels***

   If audibility is the only factor contributing to speech intelligibility then it can be assumed that hearing impaired listeners would demonstrate equal speech intelligibility ability to normal hearing listeners if the signal is presented suprathreshold. The most common method for assessing this is presenting speech in background noise at a set signal-to-noise ratio (SNR) and increasing the overall level of speech and noise signal until the listener perceives the speech correctly (Moore, 1998). At high presentation levels the speech intelligibility thresholds of hearing impaired listeners are poorer than those for normal hearing individuals (for example, Plomp, 1978; Plomp, 1986; Smoorenburg, 1992; Summers et al, 2013). The performance gap observed between normal hearing and hearing impaired individuals, listening to a suprathreshold speech signal, also leads to the conclusion that factors other than audibility contribute to speech intelligibility.

To summarise, there is no doubt that audibility plays a vital role when listening to speech; if the signal is not audible then the listener cannot access the information. However, given that hearing

impaired listeners are outperformed by normal hearing listeners, even when the speech signal is made audible, and the SII is unable to accurately predict speech intelligibility for hearing impaired listeners, it is reasonable to assume that factors other than audibility contribute towards speech intelligibility. These suprathreshold factors are addressed below. The psychoacoustic factors are covered in Section 4.2.1, and Section 4.2.2 will explore some of the non- psychoacoustic factors.

4.2.1 Psychoacoustic abilities contributing to speech intelligibility

The suggestion that factors other than audibility contribute towards speech intelligibility is not novel. By far the most influential and widely cited text on this topic is that by Plomp (1978), who noticed the limited benefit that linear hearing aids were providing when listening to speech in everyday listening environments; despite the speech being made audible there was still a noticeable performance gap between hearing impaired and normal hearing listeners when listening to speech. This prompted Plomp (1978) to propose a description of hearing impairment which presented hearing loss as being made up of two elements: 1) attenuation, Factor A, and 2) distortion, Factor D. The attenuation factor accounts for the threshold shift aspect of hearing loss which can be compensated for by increasing the level of sound pressure to the ear. The distortion factor has been introduced to address what is occurring when an individual says 'I can hear that people are talking but I can't understand them' (Plomp, 1978, p.537). Factor A can be measured using audiometry but it is the Factor D element that is not as easily quantified and causes the relationship between PTA and real world performance to be non-linear.

The purpose of Section 4.2.1 is to provide the reader with a general understanding of the psychoacoustic abilities that are not assessed by PTA but are utilised when listening to speech. In-depth coverage of the physiology and psychology of hearing and hearing impairment is not provided here; for a review of this the reader is referred to Moore's book 'An Introduction to the Psychology of Hearing' (2008a). The components that make up the distortion element and how they interact are still not fully understood (Moore, 1998; Moore, 2003; Moore, 2008a). However, it is generally agreed that there are three main suprathreshold components which contribute towards the intelligibility of speech in either a quiet or noisy environment; frequency selectivity, temporal resolution and loudness recruitment (Moore, 2008a). These three psychoacoustic abilities are covered in the following three numbered sections, focusing on how they are influenced by hearing impairment, how they contribute towards speech intelligibility, and evidence that individual variation cannot be explained by PTA.

1. Frequency selectivity

The tonotopic organisation of the basilar membrane leads to a series of overlapping auditory filters, each with a different centre frequency and bandwidth, similar to a number of band pass filters (Fletcher, 1940). At the apex of the cochlea the inner hair cells on the basilar membrane are tuned to detect low frequencies and the bandwidth of the auditory filters is broader than those at the base of the basilar membrane, where high frequency sounds are detected (Moore, 2008a). In normal listeners the auditory filter shape is similar to that shown in Figure 4.1 (left) but in an impaired ear the auditory filter has a different shape, similar to that shown in Figure 4.1 (right). The broadening and flattening of the auditory filters has a direct impact on frequency selectivity and temporal resolution (Moore, 2008a).

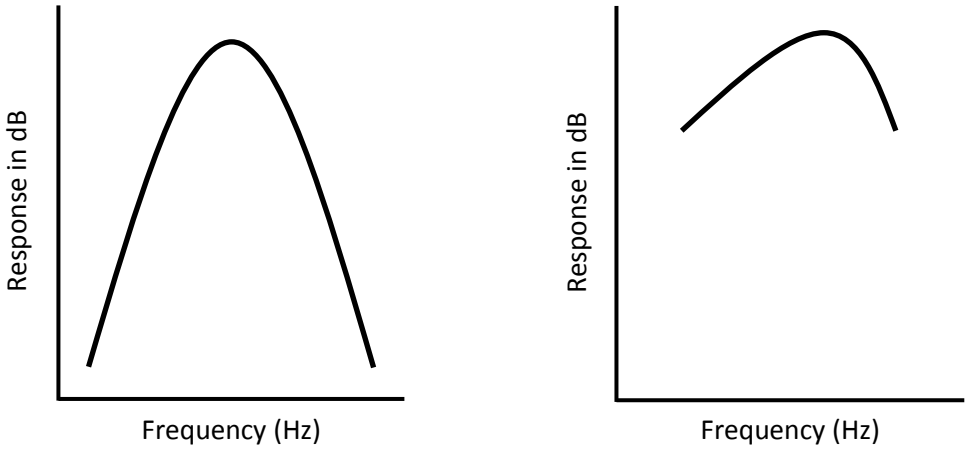


Figure 4.1 *Schematic representations of a normal hearing (left) and a hearing impaired (right) auditory filter*

Frequency selectivity can be defined as the ability to distinguish between the spectral components of complex sounds. The broadening and flattening of the auditory filters observed in impaired listeners results in a 'smeared' response of the auditory filters, as each filter is stimulated by a wider range of frequencies. According to Moore (2008a) this has two main perceptual consequences. Firstly, when listening to complex sounds, such as speech, impaired frequency selectivity reduces the ability to detect the differences in the spectral composition of a sound, which compromises distinction between different timbres. This results in it being more difficult for hearing impaired listeners to distinguish between different vowel sounds, even when the sound is made audible. Secondly, when listening to a signal in background noise the listener attends to the auditory filter with the best SNR. In a normal hearing ear the auditory filters are

relatively narrow which means that the SNR ratio within the auditory filter with a centre frequency matching that of the signal is relatively low; all of the background noise is attenuated apart from the narrow band around the signal frequency. The broader and flatter auditory filters of an impaired ear result in more of the noise entering the auditory filter with the same centre frequency as the signal, making it harder for the listener to distinguish between the signal and masker. As a result, background noise has a greater impact on the detection and discrimination of speech for hearing impaired listeners than normal hearing listeners (Moore, 2003).

It is difficult to directly evaluate the role of frequency selectivity when listening to speech because there is no way to isolate this psychoacoustic factor from other suprathreshold factors influencing performance. Baer and Moore (1994) used simulation of broadened auditory filters to measure the impact of this factor on speech intelligibility in noise, with normal hearing listeners. It is acknowledged that there are limitations to using a simulation of broadened auditory filters with normal hearing listeners to predict the performance of hearing impaired listeners, but the benefit of this method is that it allows for the effect of frequency selectivity to be assessed in isolation. They showed that in a difficult listening scenario (listening to speech with a competing speech masker, 9 dB more intense), simulated broadening of the auditory filters caused a reduction in performance. This effect was increased considerably for more adverse listening conditions (decreased SNR). Subsequently, numerous papers have used Baer and Moore's simulation to demonstrate the impact of frequency selectivity on speech intelligibility in noise (e.g. Léger et al, 2012 and Xu et al, 2012), however studies have also shown that reduced frequency selectivity alone is not able to account for all the difficulties experienced by hearing impaired listeners when listening to suprathreshold speech (Gnansia et al, 2009; Bernstein and Brungart, 2011).

2. Temporal resolution and temporal fine structure

The changes in an auditory signal over time are known as 'temporal cues' and certain information about the signal is contained within these cues. For speech, temporal fluctuations in the waveform carry important cues for speech intelligibility. Some information is contained in the long term properties of the temporal envelope (Figure 4.2, the red line), such as the prosodic information, e.g. intonation, stress and rhythm. The short-term fluctuations, known as the temporal fine structure (TFS, Figure 4.2, the blue line) contain different information, such as the segmental properties of speech, e.g. the articulation of consonants and the voicing of phonemes (Reed et al, 2009). An individual's temporal resolution ability commonly refers to detecting changes in the temporal envelope, whereas an TFS ability refers to detecting changes in the rapid fluctuations (Moon and Hong, 2014).

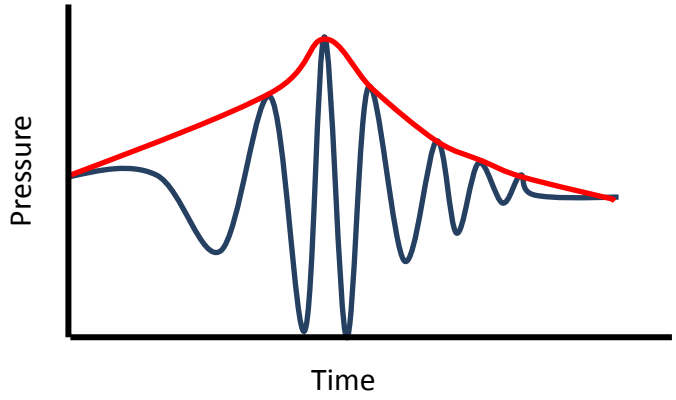


Figure 4.2 *Schematic representation of the temporal information contained in the long-term properties of the temporal envelope and short-term fluctuations of a speech signal*

According to Moore (1998) and Reed et al (2009) there are two main processes which account for the temporal processing of sounds. Firstly, within each auditory filter the time pattern of the signal is analysed. Secondly, the time pattern differences between auditory filters are compared, providing information about the temporal differences across the frequency spectrum. The envelope cues provide some important information for speech intelligibility and until recently it was widely believed that the envelope was more important than the TFS (Drullman, 1995). However, several more recent studies have concluded that the information contained in the TFS is of more importance for speech intelligibility, especially when listening in the presence of noise, and is affected to a greater extent by hearing impairment (Moon and Hong, 2014, Reed et al, 2009; Lorenzi et al, 2006; Moore, 2008b; Moore, 1998).

Hopkins et al (2008), Hopkins and Moore (2009) and Lorenzi et al (2006) have consistently reported that being able to detect TFS is necessary for speech recognition in noise. Hopkins et al (2008) assessed speech recognition in the presence of a competing talker using signals with variable amounts of TFS information. The normal hearing listeners showed a significant improvement in speech recognition scores as TFS information was re-introduced. In addition, the hearing impaired listeners did not benefit as much the normal hearing group from the additional TFS information and there was a lot more individual variation in speech recognition scores for those with hearing loss. Hopkins and Moore (2009) used the same approach as Hopkins et al (2008), measuring speech recognition in noise for signals with varying amounts of TFS information. As expected, normal hearing subjects benefited from additional TFS information and this benefit was greater when listening in fluctuating noise in comparison to steady noise. This finding suggests that TFS information is important for utilising the fluctuating changes in SNR to detect speech cues when the SNR is more advantageous, known as ‘dip listening’. Finally, Lorenzi et al (2006) processed speech signals so as to preserve either the envelope cues or the TFS cues. Normal hearing and hearing impaired listeners were presented with unprocessed speech and the

two types of processed speech in quiet. Hearing impaired listeners performed poorly when listening to the speech with TFS cues preserved. This indicates an inability to utilise the TFS cues in comparison to normal hearing listeners and highlights the importance of this psychoacoustic ability when listening to speech.

3. Cochlear compression and loudness recruitment

The phenomenon of cochlear compression allows for a normal cochlea to perceive a wide range of sounds (Moore, 2002). It is commonly accepted that the compression of the basilar membrane arises as a result of the operation of an 'active' physiological mechanism of the outer hair cells (Moore, 2008a, pp. 30 & 157). If there is a loss of function of the outer hair cells (often observed in sensorineural hearing impairment), this leads to a loss of cochlear compression (Moore, 2002). This is associated with reduced dynamic range (the range between an individual's threshold and uncomfortable loudness level) and loudness recruitment (Moore and Glasberg, 1997). Loudness recruitment is the abnormally-rapid growth in loudness perception experienced by individuals with increased pure-tone thresholds and a reduced dynamic range (Moore & Glasberg, 1993). For a hearing impaired listener, although their hearing thresholds are elevated, when sounds are presented above their threshold the rate at which loudness perception grows is greater than that observed in normal hearing listeners (Moore, 2003). The model proposed by Moore and Glasberg (1993) states that normal and impaired listeners perceive a 100 dB SPL tone at 1000Hz as the same loudness level. However, the perceived loudness level for those with impaired hearing increases more rapidly between their threshold and a 100 dB SPL presentation than it does for those with normal hearing, as shown in Figure 4.3. Reduced cochlear compression is thought to be the main cause of loudness recruitment in hearing impaired listeners (Moore, 2002), but there is still some debate about the specific mechanisms that cause this phenomenon (Joris, 2009).

Cochlear compression, as well enabling the cochlear to detect a wide range of stimulus intensities, plays additional roles in auditory perception, including intensity discrimination, detecting a signal in the presence of a competing noise (masking), loudness perception and timbre perception (Moore, 2002; Oxenham and Bacon, 2003). According to Moore (2003), reduced cochlear compression may negatively impact speech intelligibility in three ways. Firstly, when listening in a fluctuating background noise impaired listeners may demonstrate reduced 'dip listening' ability; the highest comfortable presentation level of the background noise may be only slightly higher than the individual's threshold, resulting in the speech level in the dips being close to or below threshold. Secondly, the loudness relationship between the components of the speech signal can be distorted, resulting in the relative loudness levels of the different components being different from those perceived by normal hearing listeners. Thirdly, the perception of a speech signal's

amplitude modulation can be distorted, leading to impaired listeners perceiving modulation depths to be greater than normal hearing listeners. However, there is some debate as to whether modulation depth perception is important for speech intelligibility so the impact of this aspect of cochlear compression is unclear (Moore, 2003).

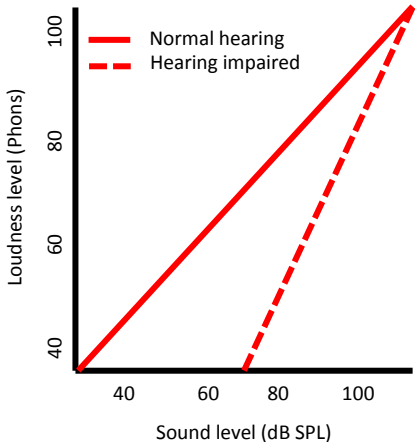


Figure 4.3 *Diagram showing the concept of loudness recruitment proposed by Moore and Glasberg (1993)*

4. Summary

Section 4.2.1 has provided an overview of the suprathreshold deficits experienced by hearing impaired listeners which influence speech intelligibility ability and are not assessed by PTA. These deficits help to explain why hearing impaired listeners are outperformed by normal hearing listeners on speech intelligibility tasks, even when the signal level is raised to be within an audible range. It is also worth mentioning hidden hearing loss here, another issue relating to suprathreshold deficits which are not detected through PTA. It has been shown that individuals with normal audiograms (hearing thresholds lower than 20 dB HL) can also display difficulty when listening to speech, in particular in noisy environments (Zhao & Stephens, 2007). This auditory deficit is particularly prevalent in populations exposed to high levels of noise (Plack et al, 2014), an issue within the military population. Plack et al (2014) stated that there is evidence of an association between a history of noise exposure and difficulties with the perception of speech, even when there is no evidence of any audiometric loss. Results from animal studies suggest that noise exposure causes damage to auditory nerve fibres that is not detected by audiometry and similar studies with humans have shown that noise exposure can be associated with suprathreshold deficits that are not detectable from the audiogram (Plack et al, 2014). This presents two challenges when considering using PTA as a measure of AFFD. Firstly, individuals with normal thresholds will be classified as being fit for duty, despite potential difficulties carrying out the SC-MCATs. Secondly, individuals who potentially have the early stages of NIHL are not

being detected through PTA and therefore not taking the necessary precautions to protect their ears from further damage.

4.2.2 Non-psychoacoustic factors influencing speech intelligibility

It has been established that PTA does not measure a number of psychoacoustic abilities that impact speech intelligibility. However, the factors that influence performance on the SC-MCATs in an operational environment cannot be limited to auditory abilities. There is a great deal of research exploring the cognitive abilities that cause individual differences when listening to speech. When people listen to spoken language they apply their knowledge of language, add context based on their understanding of the real world and use previous experiences to make predictions about what is likely come next; using this information is referred to as 'top down processing' (Akeroyd, 2008). In addition, factors such as working memory capacity can influence speech intelligibility (Heinrich et al, 2015). The influence of cognitive factors, which encompasses a wide range of non-psychoacoustic abilities, on speech intelligibility is widely researched (see Akeroyd, 2008 for a recent review of this). The non-psychoacoustic factors which may influence performance on the SC-MCATs, and are not assessed by PTA, are briefly outlined in the following paragraphs.

Knowledge of language and vocabulary (associated with job experience)

Some of the phrases used within the military are specific to the occupation. Phrases such as 'form a hasty defence' and 'two one delta stop' all contain recognisable language but would not be routinely spoken by civilians. Parallels can be drawn between this situation and individuals listening to speech not in their native language. Evidence suggests that non-native listeners are outperformed by native listeners when presented with speech in noise (Mayo et al, 1997; Na´belek and Donahue, 1984; Rogers et al, 2006). It may be the case that that military personnel with more experience and knowledge of the syntactic and semantic elements of military communication are able to use this to enhance their performance when listening to commands in adverse listening situations.

Age

Several studies have suggested that speech understanding in older people is limited not only by a decline in auditory abilities but also by cognitive decline associated with aging (e.g. Gordan-Salant & Fitzgibbons, 1997; Versfeld and Dreschler, 2002; Zekveld et al, 2011). If age impacts performance on the SC-MCATs then this factor should be considered when assessing AFFD.

Ability to cope in stressful situations

It is predicted that individual differences in the ability to cope in stressful situation will impact performance on the SC-MCATs and ultimately, indirectly, their AFFD. However, this is a prediction; no evidence has been found to demonstrate the influence of stressful situations on speech intelligibility.

Working memory

There is no universally accepted definition of working memory but it is commonly used to refer to the simultaneous storing and processing of information (Heinrich et al, 2015). It is widely accepted, although not fully understood, that there is a link between working memory and speech intelligibility (Gordan-Salant and Fitzgibbons, 1997; Heinrich et al, 2015). In relation to AFFD, two individuals with similar audiometric thresholds, but different working memory capacity, are likely to vary significantly in their performance carrying out SC-MCATs, a difference that would not be detected by PTA.

Sensory modality impairments other than hearing loss

Finally, as discussed in Section 2.2, during the information-gathering stage of situational awareness, critical factors about the surrounding environment are obtained from various sensory modalities, such as visual and tactile perceptions (Endsley, 1995). If an individual has, for example, impaired vision, they may miss important information relevant to the SC-MCATs, impacting their performance. Also, the availability of lip reading cues when listening to speech in noise is known to influence performance (evident from the McGurk effect; Nath and Beauchamp, 2012).

4.2.3 Summary

It has been explained in Sections 4.2.1 and 4.2.2 that there are both psychoacoustic and non-psychoacoustic factors, which are not assessed by PTA, that influence speech intelligibility. It is therefore assumed that these factors will also impact performance on the SC-MCATs and therefore AFFD. However, just because the results from PTA cannot *explain* the influence of these factors on speech intelligibility tasks does not necessarily lead to the conclusion that PTA cannot *predict* speech intelligibility. A method for further exploring this is covered in Section 4.3.

4.3 An approach for exploring the predictive validity of PTA as a measure of performance on the speech communication MCATs

Section 4.3 key argument
The association between PTA and speech intelligibility tests can be used to assess whether PTA is a suitable tool for predicting performance on the SC-MCATs or whether alternative assessment methods should be explored.

It has been established that PTA does not assess all of the psychoacoustic and non-psychoacoustic factors that influence performance when carrying out speech intelligibility tasks. When considering AFFD assessment, tools must be able to accurately predict individual performance on the MCATs, and of particular interest here, the SC-MCATs. Section 4.2 has outlined why, in principle, it is thought that PTA may be poor at predicting performance on the SC-MCATs. Section 4.3 outlines a framework for using the correlations between PTA and performance on speech intelligibility tests to indicate the suitability of PTA as a predictor of performance on the SC-MCATs.

Speech intelligibility is a measure of an individual's ability to process speech sounds and is commonly used to predict how well individuals perform in everyday listening scenarios. It is normally tested by presenting speech stimuli (such as single words, phonemes, or sentences) either in quiet or background noise, measuring the listener's speech recognition threshold (SRT). The exact definition of an SRT varies greatly between tests but a common measure is the presentation level (if measured in quiet) or the SNR (if measured in noise) at which a listener scores 50% correct (Schoepflin, 2012). Greater detail about the measurement of speech intelligibility is provided in Chapter 5.

Figure 4.4 explains how looking at the relationship between PTA and speech intelligibility tests will help to provide an indication as to whether PTA is a suitable tool for predicting performance on the SC-MCATs and ultimately whether it should be considered as a suitable generic predictive test (GPT, see Section 2.4.3) for assessing AFFD. One approach towards investigating this would be to develop a simulation of the SC-MCATs and to measure the association between PTA and performance on the simulation. However, this would be time-consuming and it is thought that sufficient evidence can be found by exploring the relationship between PTA and performance on speech intelligibility measures. Justification for this is provided in the following paragraph; the bracketed numbers refer to the six points in Figure 4.4.

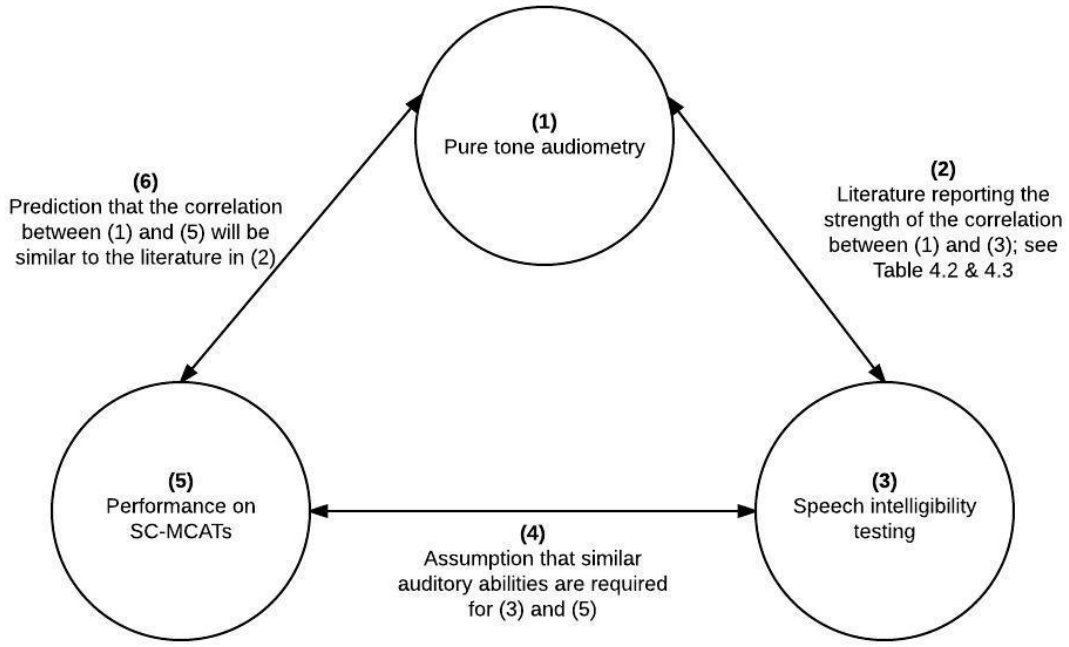


Figure 4.4 *Diagram explaining how the literature reporting the correlations between PTA and speech intelligibility testing can be used to predict the suitability of PTA as a measure of AFFD*

There is evidence in the literature (2) that reports the relationship between PTA (1) and speech intelligibility tests (3), outlined in tables 4.1 and 4.2. An assumption (4) has been made that the auditory abilities being tested when measuring speech intelligibility (3) will be similar to those required to carry out the SC-MCATs (5). Based on this assumption (4), it is possible to predict that the correlation (6) observed between PTA (1) and performance on the SC-MCATs (5) would be similar to that reported in the literature (2) between PTA (1) and speech intelligibility tests (3).It therefore follows that if a strong correlation is observed between PTA and speech intelligibility tests it is possible that PTA may be a suitable tool for predicting performance on the SC-MCATs. However, if a weak correlation is observed or the literature is inconclusive about the relationship between (1) and (3) this is indicative of PTA not being a suitable tool for predicting performance on the SC-MCATs and is justification for exploring alternative assessment tools.

It is acknowledged that there are limitations to the argument outlined by Figure 4.4. Firstly, some of the non-psychoacoustic factors that might influence performance on the SC-MCATs (outlined in Section 4.2.2) will not be assessed by speech intelligibility tests. For example, when carrying out SC-MCATs in an operational scenario, individuals may be under a lot of stress or they may have other tasks competing for their attention; this ability is not measured by a speech intelligibility test. Since it is expected that PTA is also not measuring these factors it is a reasonable assumption that the levels of correlation observed in point two of Figure 4.4 may be lower at point six. Secondly, there is a great deal of methodological variation in the studies reported in Table 4.2

(such as the PTA predictive configuration and the choice of speech intelligibility test) making it difficult to select a single correlation value which represents point two of Figure 4.4. Nonetheless, assessing the literature that reports the correlations between point one and three of Figure 4.4 will provide an indication as to whether there is scope for using PTA to predict performance on the SC-MCATs.

Most commonly, studies report the association between PTA and speech intelligibility tests as a correlation coefficient. The correlation coefficient between two sets of data is a way of expressing how closely the two data sets are related. The most common method for assessing this (with normally-distributed data) is the Pearson Product Moment Correlation, or more commonly named the Pearson Correlation. Correlation values are expressed as a coefficient, ranging from -1 to +1. A negative correlation coefficient indicates that as one variable increases the other decreases. A positive correlation coefficient indicates that as one variable increases the other also increases. A correlation coefficient of zero indicates there is no linear relationship between the two variables (Field, 2005a). It is known that in general there is a positive correlation between PTA and speech intelligibility; as hearing threshold levels worsen speech intelligibility scores also worsen (see Table 4.1 for evidence of this). The correlation coefficient value squared provides an estimation of how much variance is shared between the two variables being measured (Khan Academy, 2015). In the context of Chapter 4, this refers to how much variation in the speech intelligibility tests is shared with PTA. An assumption is being made that measures of speech intelligibility are assessing similar auditory abilities required to carry out the SC-MCATs identified in Chapter 3. Therefore, put simply, if the literature reports strong correlations between PTA and measures of speech intelligibility this is evidence to suggest that PTA may be a suitable tool for predicting performance on the MCATs described in Chapter 3. If the literature reports weak correlations this indicates that a large amount of the variation observed in measures of hearing ability is not shared with PTA, suggesting it may not be a suitable tool for measuring AFFD.

Since an individual's audiogram is made up numerous pure-tone thresholds, quantifying this as a single value (as is required for correlation calculations) is a challenge. King et al (1992) conducted a review of different audiometric predictive configurations and concluded that the evidence does not advocate any single descriptor as a superior predictor over another. An enormous amount of literature is available assessing the best pure-tone configuration for predicting overall hearing ability, measured either by self-report questionnaires or speech intelligibility tests. The most commonly cited of these is Fletcher (1950), despite being written over 60 years ago. Fletcher (1950), developed one of the first formulae for calculating speech intelligibility ability from the audiogram and proposed examining the thresholds at 500, 1000 and 2000 Hz and using the

average of the two smallest values to predict the audibility of speech at threshold level. Numerous subsequent papers have proposed alternative predictive configurations, with a variety of different hearing ability measures being used and a general consensus for a single best formula has not been met (Andrade et al, 2013). Although it is worth acknowledging the difficulty of describing the audiogram as a single figure, the purpose of this Chapter is not to evaluate what the best audiometric configuration is for predicting speech intelligibility. In Section 4.4 literature looking at the association between PTA and measures of hearing ability will be explored, regardless of which PTA configuration has been used.

4.4 Reporting the association between PTA and speech intelligibility tests

Section 4.4 key arguments
PTA is unable to account for (in a statistical sense) an adequate amount of the variation in performance in speech intelligibility tests to justify solely investigating audiometric thresholds to assess AFFD. *In comparison to PTA, speech intelligibility tests may be better able to account for the factors that cause variation between individuals on the SC-MCATs.*

It has been established that the literature reporting correlation coefficients between PTA and speech intelligibility tests can be used to assess the suitability of PTA for predicting performance on the SC-MCATs. In Section 4.4 a number of studies which report the correlation between these two measures are outlined. The papers reported in tables 4.1 and 4.2 were selected following a literature search using PubMed, Google Scholar, Web of Science and DelphiS which aimed to find papers reporting the relationship between PTA and speech intelligibility tests. A combination of the following search terms was used: hearing disability; pure-tone audiometry; audiometry; speech recognition; speech in noise/quiet; speech intelligibility; speech perception. In addition the reference lists of identified relevant papers and text books (including King et al, 1992) were searched for any papers missed by the online search. This search identified a large number of papers (over thirty) so a decision was made to focus on a selection of papers which had a clear methodology and reported the relationship between PTA and measures of hearing ability as a correlation coefficient, for ease of reporting and comparing the results. In addition, any papers with small sample sizes (<10) or a population with a narrow spread in hearing acuities were disregarded due to the negative impact of these factors on calculating a correlation coefficient. It should be acknowledged that a systematic review of all the literature in this area would be beneficial but has not been conducted within the scope of this thesis. The papers reported here do not constitute a full review of all the literature but simply those identified using the criteria listed above. Although the author did not consciously select papers based on the correlation

values reported, it is important to acknowledge that there is potential bias in those selected since the author may be subconsciously biased towards choosing papers which displayed poor correlations, supporting the argument that PTA is not fit for duty.

It is proposed that a proper systematic review (a method for summarising research evidence; Hemingway and Brereton, 2009) is completed, aiming to answer the question, 'to what extent can audiometric thresholds be used to predict speech intelligibility in quiet and in noise?' This information would be useful for the evaluation of the suitability of PTA, not only as a tool for assessing AFFD, but also as a tool for assessing hearing ability in audiology clinics. A systematic review could be followed up by a meta-analysis, a method for combining the results from multiple studies to provide a statistical estimate of an effect (Crombie and Davies, 2009).

Table 4.1 contains 40 correlation coefficient values which vary from 0.50 to 0.95, taken from nine sets of experimental data. This indicates that anything from 25% to 90% of the variation observed in the speech intelligibility tests is shared with the variation observed in PTA. For every correlation reported, the 95% confidence interval has been calculated using the equation outlined in Altman et al (2000, p.89), detailed in Equations 1-5. According to Altman et al (2000, p.91) the distribution of Spearman's r_s is similar to Pearson's r, so confidence intervals for both correlation values can be can be calculated using the same method. First the correlation r value is transformed to a quantity Z, which has an approximately normal distribution (Equation 1). Then F and G quantities are calculated (Equations 2 and 3) which are then transformed back to the original scale to provide the upper (Equation 4) and lower (Equation 5) confidence limits. Taking the 95% confidence intervals of the correlations coefficients reported into account, it could be that as little as 3% or as much as 96% of the variation observed in the speech intelligibility tests is shared with the variation observed in PTA.

Equation 1

$$Z = \frac{1}{2}\log e \left(\frac{1+r}{1-r}\right)$$

Equation 2

$$F = Z - \frac{1.96}{\sqrt{n}}$$

Equation 3

$$G = Z + \frac{1.96}{\sqrt{n}}$$

Equation 4

$$lower\ 95\%\ CI = \frac{e^{2*F} - 1}{e^{2*F} + 1}$$

Equation 5

$$upper\ 95\%\ CI = \frac{e^{2*G} - 1}{e^{2*G} + 1}$$

The majority of the correlation coefficients reported in Table 4.1 lie between 0.6 and 0.75; 23 of the 40 values cited are within this range. The correlation values observed between audiometric and speech-in-quiet tests are on average slightly better than those observed with SIN tests; the

average correlation across 15 values between PTA and speech-in-quiet tests is 0.71 (50% shared variance) compared to 0.69 (48% shared variance) between PTA and SIN tests, averaged across 25 values. Considering that the majority of the SC-MCATs involve listening to speech in the presence of interfering noise it is fair to assume that the slightly lower correlation observed between PTA and SIN tests is more likely to represent what would be observed between PTA and the SC-MCATs.

Four statements can be given based on that the data reported in Table 4.1:

1. Overall, there is a positive correlation between PTA and speech intelligibility tests; PTA is able to predict that, in general, as audiometric thresholds deteriorate performance on speech intelligibility tests also worsens but it is not able to explain why two individuals with very similar audiometric data can display very different scores on a speech intelligibility test.
2. There is no clear agreement about the correlation strength between PTA and measures of speech intelligibility.
3. PTA is unable to account for all of the observed variation in speech intelligibility tests.
4. It is predicted that PTA will be unable to account for all the variation observed in performance on the SC-MCATs and therefore will be a poor predictor of performance on these tasks.

The mean of the 40 correlation coefficients reported is 0.69 (0.57-0.78, mean of lower and upper 95% confidence intervals); this indicates that, averaged across all methodological differences and PTA predictive configurations, PTA is on average only able to account for less than half (48%) of the observed variation in speech intelligibility tests. This result does not support the use of PTA as a GPT for predicting performance on the SC-MCATs, especially considering that it is expected that the association between PTA and performance on the SC-MCATs will be weaker than that observed between PTA and speech intelligibility tests (see Section 4.3).

The main conclusion that can be drawn from Table 4.1 is that individuals with very similar audiometric results can display very different performance on speech intelligibility tests, resulting in the correlation between the two variables being weak. In summary, the best method for assessing an individual's speech intelligibility ability is to directly measure it, rather than using the audiogram as a predictive tool. Based on this evidence, it is proposed that a measure of speech intelligibility ability is considered as an alternative measure of AFFD for assessing individual performance on the SC-MCATs; this is addressed in Chapter 5.

Table 4.1 *Literature reporting correlations between PTA and speech intelligibility tests in quiet (SQ) or noise (SN) with normal hearing (NH) and hearing impaired (HI) listeners*

Study	Sample	Speech in quiet or noise	Correlation (*r*) between PTA (across various configurations, kHz) and speech intelligibility measures (see Table 4.2 for details of study) with 95% confidence interval for *r* calculated using the equation outlined by Altman et al (2000, p.89)									
			0.5,1,2	**0.5,1,2,3**	**0.5,1,2,4**	**1,2,3**	**1,2,4**	**1,2,3,5**	**2,3**	**2,4**	**2,3,4**	**3,4,6**
1) Harris, 1965	52 HI	SQ	0.55 (0.32-0.71)			0.74 (0.58-0.84)						
2) Humes and Roberts, 1990	13 HI	SQ					0.91 (0.72-0.97)					
		SN					0.90 (0.69-0.97)					
3) Smoorenburg, 1992	200 HI	SQ	0.73 (0.66-0.79)	0.68 (0.60-0.75)	0.68 (0.60-0.75)	0.63 (0.54-0.71)		0.58 (0.48-0.67)	0.57 (0.47-0.66)	0.56 (0.46-0.65)	0.55 (0.45-0.64)	
		SN	0.50 (0.39-0.60)	0.63 (0.54-0.71)	0.67 (0.59-0.74)	0.66 (0.57-0.73)		0.72 (0.65-0.78)	0.68 (0.60-0.75)	0.72 (0.65-0.78)	0.72 (0.65-0.78)	
4) Kramer et al, 1996	51 HI	SQ			0.61 (0.40-0.76)							
		SN			0.53 (0.30-0.70)							
5) Peters et al, 1998	27 11=NH 16=HI	SQ	0.95 (0.89-0.98)				0.94 (0.87-0.97)			0.93 (0.85-0.97)		
		SN	0.52 (0.17-0.75)				0.58 (0.26-0.79)			0.61 (0.30-0.80)		
6) Leensen, 2013- Dutch Sentence Test	98 49=NH 49=HI	SN			0.82 (0.74-0.88)					0.82 (0.74-0.88)		0.80 (0.72-0.86)
7) Leensen, 2013- National Hearing Test		SN			0.72 (0.61-0.80)					0.74 (0.63-0.82)		0.69 (0.57-0.78)
8) Leensen, 2013- Earcheck		SN			0.66 (0.53-0.76)					0.64 (0.51-0.74)		0.62 (0.48-0.73)
9) Leensen, 2013- Occupational Earcheck		SN			0.69 (0.57-0.78)					0.67 (0.54-0.77)		0.66 (0.53-0.76)

Table 4.2 *Overview of methods used in papers listed in Table 4.1*

Speech intelligibility measure descriptions

1) Sentences from the Silverman and Hirsh (no further reference given) were used. Speech intelligibility was measured as the 'Discrimination Score' which is the percentage of words correctly identified at 40dB above the individual's SRT.

2) Sentences from the City University of New York (CUNY) sentence test were presented in quiet and noise. Speech intelligibility was measured as the percentage correct scores at a set presentation level.

3) Sentences from the Dutch sentence SRT test were presented in quiet and speech-spectrum noise. An adaptive procedure was used; the speech level was varied and for the noise condition the noise level was fixed. Thirteen sentences were presented; the average presentation levels of the final ten sentences were used to calculate the speech intelligibility score.

4) Dutch sentences (specific test not named) spoken by a female speaker were presented in both quiet and steady-state noise. An adaptive procedure was used; varied speech and for the noise condition the noise level was fixed. Thirteen sentences were presented; the average presentation levels of the final 10 sentences were used to calculate the speech intelligibility score.

5) Sentences from the Hearing in Noise Test (HINT) were presented in quiet and in speech-spectrum noise. An adaptive procedure was used; varied speech and for the noise condition the noise level was fixed. Details of the adaptive procedure method are not provided.

6) The same method employed by Smoorenburg (1992) was used here (see above, point 3).

7) Dutch monosyllabic digits were used to create a set of 80 digit triplets which were presented in speech-spectrum noise. An adaptive procedure (fixed noise and varied speech level) was used; Twenty-three words were presented; the average presentation levels of the final 20 triplets were used to calculate the speech intelligibility score.

8) Dutch consonant-vowel-consonant (CVC) words, selected from the Dutch wordlist used for diagnostic speech audiometry, which are phonemically balanced and representative of the Dutch language, were presented in speech-spectrum noise. An adaptive procedure (fixed noise and varied speech level) was used. Twenty-seven words were presented; the average presentation levels of the final 20 CVC words were used to calculate the speech intelligibility score.

9) The same stimuli as used in (7) but only including words which contain a high proportion of high frequency consonants, presented in speech-spectrum noise. An adaptive procedure (fixed noise and varied speech level) was used. Thirty-five words were presented; the average presentation levels of the final 30 CVC words were used to calculate the speech intelligibility score.

4.5 Chapter 4 Summary

Section 4.5 Concluding statements
There is evidence to suggest that PTA may not be a good tool for predicting performance on the SC-MCATs. ***A speech intelligibility test should be investigated as an alternative tool for predicting performance on the SC-MCATs.***

The focus of this chapter has been to investigate the suitability of PTA as a tool for predicting performance on the SC-MCATs and ultimately as a tool for assessing AFFD. This has been explored in two ways. Firstly, the theoretical argument that PTA, in principle, cannot explain individual differences in performance on speech intelligibility tests has been outlined. It has been discussed that there are psychoacoustic and non- psychoacoustic factors that influence speech intelligibility which are not assessed by PTA. Secondly, the correction between audiometric thresholds and performance on speech intelligibility tasks has been explored. From the literature reviewed, it is clear that there is no clear agreement about the correlation strength between PTA and measures of speech intelligibility, and PTA is unable to account for all of the observed variation in speech intelligibility tests. It is acknowledged that a proper meta-analysis should be conducted in order to gather a more accurate picture about this relationship than is reported here (see Section 7.3, suggestions for future work). It is assumed that an equally weak, if not weaker, correlation will be observed between PTA and performance on the SC-MCATs as was reported with speech intelligibility tests. The added influence of the non-psychoacoustic factors (see Section 4.2.2), affecting performance on the SC-MCATs, is likely to further weaken this correlation. It is proposed that speech intelligibility testing is explored as an alternative method for predicting performance on the SC-MCATs. The focus will be on speech-in-noise tests as the majority of the SC-MCATs involve listening to speech in the presence of a noise masker rather than listening to speech-in-quiet. Chapter 5 provides an overview of SIN tests and the process for selecting, developing and validating one which may be suitable for use as a measure of AFFD.

Chapter 5: Developing a new auditory fitness for duty measure

5.1 Introduction

The MCATs carried out by infantry and combat support personnel were identified in Chapter 3. Seven out of the ten MCATs which should be prioritised for representation in a measure of AFFD for infantry and combat support personnel are speech communication tasks. Chapter 4 has explained why it is thought that PTA may not be able to accurately predict performance on these SC-MCATs. Chapter 4 has also justified why exploring the introduction of a new AFFD test which is focused on measuring speech communication ability is an appropriate next step. Chapter 5 explores the topic of speech-in-noise (SIN) testing and goes through the process for reviewing, selecting and developing a SIN measure which can be considered for implementation within a military AFFD test battery. The following topics are covered in Chapter 5:

- An introduction to SIN testing and an overview of some of the available SIN test (Sections 5.2.1 and 5.2.2)
- The important considerations when selecting a SIN test to further investigate as a measure of AFFD (Section 5.2.3)
- A pilot study investigating the face validity of different speech tests in terms of relevance to military communications (Section 5.2.4)
- Justification for why the Coordinate Response Measure (CRM) SIN test has been selected as the measure to further investigate for implementation as part of military AFFD testing (Section 5.2.5)
- Study 2. Developing and recording the CRM test stimuli (Section 5.3)
- Study 3. Equalising the intelligibility of the CRM test material in noise (Section 5.4)
- Study 4. Exploring the measurement precision of the CRM implemented in an adaptive procedure (Section 5.5)

5.2 Speech-in-noise testing: reviewing and selecting an appropriate test

5.2.1 Introduction

Speech-in-noise (SIN) testing involves presenting speech stimuli to a listener, in the presence of background noise, at a known SNR, and measuring their response to the stimuli. In the audiology clinic this type of test is often used to gauge the level of communication difficulty an individual is experiencing as a result of hearing loss or to measure the improvement in communication ability provided by a hearing aid.

For this project the focus is on using SIN testing to predict performance on the SC-MCATs identified in Chapter 3. Measuring speech recognition in noise is of greater interest for this project than presenting speech in quiet. When reviewing the list of SC-MCATs outlined in Chapter 3 it can be assumed that the majority of these tasks are carried out in the presence of background noise, whether that is radio noise, engine noise or weapon systems being fired. In addition, the results from Study 1, Part A (Chapter 3, Section 3.3) listed one of the reasons for reduced performance as 'background noise'.

The focus of Section 5.2 is to provide an overview of some of the available SIN tests and to go through the factors to consider when selecting a SIN test to further investigate as a measure of AFFD (in particular performance on the SC-MCATs). Section 5.2 is not a comprehensive overview of all aspects of SIN testing; instead it highlights the most important points to explore in terms of military AFFD testing.

5.2.2 Overview of available speech-in-noise tests

There are a large number of speech tests available, in many different languages and in a variety of formats. In order to make an informed decision about which speech test to investigate further, as a measure of AFFD for military personnel, an overview of available speech tests has been conducted. Table 5.1 is not a comprehensive list of all speech tests, but aims to provide an overview of the different types of speech test available and the types of stimuli they contain. Speech tests for which information was readily available have been included. The focus here is on a test where by the speech is presented in a noise masker, based on the results from the focus groups outlined in Chapter 3 (Bevis and Semeraro et al, 2014), in which the participants report that they are rarely in a quiet environment.

Table 5.1 *Overview of a selection of available speech tests*

Speech Test (number refers to reference at bottom of table)	Stimuli type	Example stimuli	Background masker (typically used)	Available in British English	Target age
Hearing in noise test (HINT)[1]	Sentences. Syntactically and semantically correct Whole sentence to be scored	*The wife helped her husband*	White noise spectrally matched to sentence material	No (American English)	Adult & paediatric
Bamford- Kowel-Bamford (BKB)[2]		*The clown had a funny face*	Eight-talker babble	Yes	Paediatric
Speech-in-noise (SIN) Test and Quick SIN[3]		*The lake sparkled in the red hot sun*	Four-talker babble	Yes	Adult
Coordinate Response Measure (CRM)[4]	Sentences Syntactically and semantically correct Target words to be scored, said within a carrier phrase	*Ready tiger, go to red three now.*	No set noise type	Yes	Adult
Connected Speech Test[5]	Sentences within a conversational passage (syntactically and semantically correct)	*Could not be found*	Multitalker babble (number of speakers unknown)	Unknown	Adult
Speech Recognition in Noise Test (SpRINT)[6]	Words	*Raise Door Tip Sure*	Six-talker babble	Unknown	Adult & paediatric
Modified Rhyme Test (MRT)[7]		*You will mark went please. (Options given: went, sent, bent, dent, tent, rent)*	Unknown	Unknown	Adult & paediatric
Words in Noise Test (WIN)[8]		*Food Pain Late Dodge*	Multitalker babble (number of speakers unknown)	Yes	Adult & paediatric
Four Alternative Auditory Feature (FAAF) test[9]		*Bag Back Bad Bat*	Unknown	Yes	Adult & paediatric
Automated Toy Test (ATT)[10]		*Duck Cup Plane Plate*	White noise spectrally matched to sentence material	Yes	Paediatric
AB Short Word List[11]		*Fish Duck Gap Cheese*	Unknown	Yes	Paediatric
Triple Digit Test (TDT)[12]	Three digits	*The numbers two five nine*	White noise spectrally matched to sentence material	Yes	Adult

Main reference for each listed test: *(1) Nilsson et al, 1994; (2) Bench et al, 1979; (3) Taylor, 2003; (4) Kitterick et al, 2010; (5) Cox et al, 1987; (6) Wilson & Cates, 2008; (7) Meyer Sound, n.d.; (8) Wilson & Cates, 2008; (9) Foster & Haggard, 1987; (10) Summerfield et al, 1994; (11) Boothroyd, 1968; (12) Lutman et al, 2006.*

5.2.3 Considerations when selecting a speech-in-noise test to measure military AFFD

Section 5.2.2 has provided an overview of a selection of SIN tests. Section 5.2.3 explores the important factors to consider when selecting a SIN test for use in predicting performance on SC-MCATs to assess military AFFD. For each factor (labelled A-F) a brief introduction and definition is provided, following by a discussion about its importance in terms of military AFFD testing. The aim of this section is to ensure that it is clear what the priorities are when selecting a SIN test to be used as an AFFD test.

1. Speech material

Different speech tests employ different stimuli types. These range from meaningful sentences to meaningless sentences and monosyllabic words, nonsense syllables (e.g. vowel-consonant-vowel stimuli) and digits. The type of stimuli chosen depends upon the specific application of the speech test. In particular, the type of speech material chosen has a direct impact on what auditory abilities are being measured (HearCom, 2006).

In order to develop a SIN test which is primarily testing an individual's SIN ability it was decided that the speech stimuli should include no syntactic or semantic cues. This means the individual's vocabulary or cognitive ability will have limited impact on their score. In order to ensure that an individual's experience, memory and knowledge and understanding of vocabulary and sentence structure does not impact their performance on a speech test certain types of stimuli should be eliminated. These include meaningful sentences, sentences containing a lot of information which needs to be memorised or sentences which may contain unfamiliar vocabulary. However, in order to measure speech understanding it would be advantageous for the stimuli to contain the natural dynamics of speech, such as word stress, co-articulation and dynamic range. Meaningless sentences would achieve this goal, without introducing problems associated with measuring vocabulary or cognitive skills.

Since the ultimate goal is to design a speech test which could be implemented across the UK Armed Forces and potentially as a quick to run screening measure it is important that the test can be run without the need for a trained experimenter present to score responses. For this reason a closed response format which allows for the listener to mark their response using an automated system (e.g. selecting a response on a screen) is favourable (HearCom, 2006).

It is also important that the stimuli hold high face validity when compared to command structure. The importance of this in terms of measuring AFFD is further explored in Section 5.2.3 (6). When considering the type of speech stimuli which are similar to command structure, certain types can be immediately eliminated. For example nonsense syllables, such as vowel-consonant-vowel

stimuli do not resemble the type of communication military personnel perform and the sentences from a children's speech test (such as the BKB sentences) do not have any contextual similarities with commands.

2. Speaker

Two factors to consider when selecting a speaker for a SIN test are gender and dialect. A speaker's gender affects the fundamental frequency (F0) of their voice. Men typically have a lower F0 than women because they have longer and thicker vocal cords which vibrate slower, causing a lower voice pitch (Baken, 1987). It seems sensible to assume that for an individual with a hearing loss, the F0 of a speaker impacts how much of the speech signal is within their audible range. It could therefore be predicted that an individual with a high frequency hearing loss, typical of NIHL, would struggle to hear a female speaker in comparison to a male speaker. It was not however possible to find any evidence to support this statement so it remains and assumption. However, when designing a measure of military AFFD the speech communication task should represent the job in question (Tufts, 2011). Considering that 90.2% of the Armed Forces are male (Berman and Rutherford, 2014) it can be assumed that the majority of speech communication within the military will be delivered by a male speaker. For this reason, it can be concluded that using a SIN test which has a male speaker will hold higher face validity and will be more representative of the SC-MCATs. The second factor to consider when selecting a speaker relates to dialect, which affects features of speech such as pronunciation, vocabulary and grammar (Encyclopaedia Britannica, 2015). Strong dialects can be difficult to understand by those with alternative dialects and as such should be avoided when considering a SIN test which can be implemented across the UK. It has been decided that the speaker used for a British military AFFD SIN test should have an accent similar to 'received pronunciation' which is regarded as the "standard accent of English as spoken in the south of England" and is an instantly recognisable accent (British Library, 2014). Using this accent will avoid creating a test which is only relevant for certain populations.

3. Background noise

Speech testing can either be carried out in quiet or in the presence of a noise masker. The focus here will be SIN testing, based on the results from the focus groups outlined in Chapter 3 (Bevis and Semeraro et al, 2014), in which the participants report that they are rarely in a quiet environment.

In a general context 'noise' refers to any unwanted sounds (Moore, 2008a). In the context of SIN testing, the term noise is used to describe a signal which interferes with the speech signal being presented. The noise can be referred to as a masker; a masker is a sound which when presented

simultaneously with another signal raises the threshold of the signal (Moore, 2008a). When presenting speech in the presence of a noise masker the threshold of audibility for the speech is raised by the presence of the noise. Often the terms 'noise' and 'masker' are used interchangeable, as will be the case in this thesis.

The choice of background noise employed during a speech test can have a significant impact on the measurement results. The use of different maskers in SIN tests requires the listener to utilise different auditory processes to understand the speech, thus allowing the tester to gather information about the individuals listening ability in varying environments. For example, presenting speech in a fluctuating noise means there are gaps in the noise which gives periods of a more advantageous SNR for the listener to utilise (known as dip listening) whereas this is not possible for stationary noise. Listening to speech in this sort of masker involves utilising to spectral and temporal cues, something that hearing impaired listener's find more difficult than those with normal hearing (Christiansen & Dau, 2012). A single talker masker introduces competition for the listener's attention, since the listener is able to understand both the speaker and the masker. This type of masking can be categorised as informational masking (Brungart, 2001b) and depends on the listeners central processes to distinguish between the speech and masker (Rosen et al, 2013). These two examples simply demonstrate that different masker types can influence what auditory skills a listener needs to employ in order to listen to the speech. This is not a comprehensive overview of different masker types or of the different auditory processes involved when listening to speech in the presence of noise. Military personnel are required to listen to speech in a wide variety of noisy environments, such as in vehicle noise, radio interference, weapon systems firing or competing talkers over a radio (Study 1 part A and Bevis et al, 2014) which all present different auditory challenges.

Most commonly SIN tests use stationary speech-spectrum noise since this type of noise is not specific to any 'real world' listening environment and can be used to provide a generalised measurement of SIN ability. It also involves limited auditory processing skills making it easier to interpret what is being measured during testing. Stationary speech-spectrum noise is simply white noise that does not vary in amplitude over time and has the same frequency spectrum a speech signal. This type of noise is used for a number of speech tests, making it possible to compare performance between the different methods; this can be important during the validation stage of a new test (see text on concurrent validity in Section 5.5.1). For these reasons during the initial developmental stage of designing a SIN test to assess AFFD the background noise choice will be stationary speech-spectrum noise.

4. Presentation methods

Speech testing can either be carried out over headphones, in the sound field or in a virtual environment (such as using simulation equipment which may use varying presentation methods). If headphones are being used there is also the choice between binaural or monaural presentation.

When considering designing a speech test to measure AFFD it is important to bear in mind the environment in which they test will be run. To increase the chances of the test being implemented it should be as simple to carry out as the current test, PTA. This means it must: 1) be simple to run by trained but not specialist staff; 2) be possible to carry out in a classroom type environment; and 3) not require a large amount of specialist equipment. For these reasons the use of sound field or virtual environment presentation can be ruled out. Setting up equipment for testing in these environments can be complex. Onus is placed on the tester to control factors such as speaker distance, head movements and ambient noise levels which will all heavily impact test results. It would be very time consuming and expensive to train large number of personnel to be able to carry out reliable measurements using this method.

When presenting speech over headphones a decision needs to be made about whether to use binaural or monaural presentations methods. Monaural testing allows for the SRT of each ear to be measured separately, providing a more detailed picture of the individual's hearing ability. Within the military population there is a high incidence of NIHL, which often does not cause equal damage to both ears, causing asymmetric hearing losses. In terms of designing a speech test which is measuring the auditory skills required to carry out the MCATs identified in Chapter 3 there is reason to argue that monaural presentation would be an appropriate testing method; several of the SC-MCATs (See Chapter 3, Table 3.10) are carried out listening monaurally over a radio. In addition, many of the speech communication tasks may require personnel to initially locate a talker to gather more information, such as T1, hearing commands in a casualty situation. Monaural testing methods may allow for better predictions to be made about performance on some of the SC-MCATs.

Other the other hand, at this initial stage of AFFD test development the aim is to create a tool which will be able to quickly provide an overall indication of an individual's speech communication ability and binaural presentation is the most efficient way to do this. It is known that speech intelligibility improves when subject's listen binaurally (Persson et al, 2001) so it is important to consider that this form of testing will reveal an individual's 'best case scenario' speech intelligibility. At this initial stage of designing a SIN test to assess AFFD the stimuli will be presented binaurally.

5. Measurement procedures

To measure speech intelligibility a psychophysical procedure must be carried out. Psychophysical procedures simply refer to experimental methods which explore the relationship between physical stimuli and their subjective responses (Kingdom & Prins, 2010). There are two main procedures used when measuring speech intelligibility, the method of constant stimuli and adaptive procedures. Both of these methods gather data that can then be fitted to a psychometric function (PFs). A PF provides information about the relationship between a stimulus and the subjective responses. In relation to SIN testing, the stimulus is the SNR at which the stimuli are presented (plotted on the x axis) and the subjective response is the percentage correct at any given SNR (plotted on the y axis) and a sigmoid curve ("S" shaped) is fitted to the data points. A SRT can then be read from the PF at a specified response level. Most commonly the presentation level at which the individual scores 50% correct is taken as the 'SRT 50' score (Schoepflin, 2012). Detailed information about PFs is provided in Appendix B.

The method of constant stimuli is a psychophysical experimental method whereby stimuli are presented at a fixed level (a fixed SNR for speech-on-noise testing) either in blocks or across a range of presentations levels in a random order. The proportion of correct responses at each presentation level is scored, shown as a percentage. The range of SNRs is typically selected to ensure that data points are collected from the chance level to near 100% correct and there is the same number of presentations is given at each SNR. The choice of stimulus presentation levels is usually decided following pilot work (Kingdom & Prins, 2010). This method gathers data across the whole range of the PF, from chance level to near 100% which results in a good estimate of the slope of the PF. However, because the stimulus presentation levels are fixed, the data gathered is not focused around the location point of the PF, which can result in a poor estimate of this value. This method is favourable if it is equally important to measure the subjective responses across a range of SNRs which elicit responses from chance level to near 100%, for example when comparing PFs for different SIN tests or target words.

Adaptive procedures require fewer presentations to gather information about an individual's SRT compared to the method of constant stimuli. This is because the SNR of each trial is dictated by the response on the previous trial; rather than presenting an equal number of trials at each SNR the presentations are focused around the individual's threshold. This allows for the relevant information about their SRT to be rapidly extracted (Leek, 2001). If a correct response is obtained the SNR is lowered and vice versa, limiting the number of presentations at chance level or near 100%. This method is favouredle when measuring a threshold is of interest, since the presentation levels are focused around this point and therefore a more accurate estimation is made. This

method is not suitable if detailed information is required about the slope of the PF as limited response data is collected at the extreme ends of the PF, resulting in poorer estimates of the overall shape of the PF.

An adaptive procedure would be the most suitable psychophysical measurement procedure for running an AFFD SIN test. It is the location aspect of the PF that is most important to accurately predict, since this is the score that will be used to describe an individual's SIN ability. An adaptive procedure provides an accurate estimate of this aspect (Kingdom & Prins, 2010). The AFFD test also needs to be quick to conduct. In order to increase the chance of the test being implemented it is preferable for the test to take the same or less time to carry out than PTA; it will be harder to persuade people to introduce a new test which takes a long time to conduct, regardless of how good the test is.

6. Face validity

Face validity can be described as "whether a test 'looks valid' to examiners who take it, the administrative personnel who decide on its use and other technically untrained observers" (Anastasi, 1988, p.144). For certain types of measurement face validity can be of particular importance; one such area of testing is diagnostic tools. The results from diagnostic testing are used to conclude something about an individual, whether that is the presence of disease or something about their functional ability. It is much easier to explain test results to a patient if there is logical link between the test that was carried out and conclusions that being drawn from the test.

It is important to note that prior to considering the face validity of a diagnostic tool it should be confirmed that it has other characteristics of a good test, such as high sensitivity and specificity, measurement precision and validity (see Section 5.4 for more details on this).

During some of the focus groups conducted in Study 1 Part A participants raised the topic of their opinion about the current hearing test. If this topic was raised by participants they were asked 'What is your opinion of the current test?' and/or 'How do you feel the current test could be improved?' These questions were asked in a total of eight of the 16 focus groups. They were not put to all the groups because the question did not directly link with the motivation to explore auditory tasks carried out by infantry and combat support personnel. For this reason the responses to these questions were not included in the data analysis. Table 5.2 lists the quotes in response to questions about the current hearing test.

Table 5.2 *Quotes from Study 1 Part A. Infantry and combat support personnel's opinions on pure-tone audiometry*

Infantry personnel opinions on the current AFFD test (Bevis et al, 2014)
Because it is not 360 it doesn't represent the types of sounds being heard
If it was surround sound it would be better
Use of noises that are common in a military environment would be an improvement
Use of shouted commands would be an improvement
Something scenario based would be an improvement
Background noises should be similar to those in an military environment
The majority of servicemen are men, this should be represented in the test
The use of ear pieces similar to those used in theatre
Listening for military specific noises rather than beeps
Testing whilst using the hearing protection they use
Voice recognition over background sounds
Something based more on what you actually do out on the ground
Listening to words of commands over loud noises which replicate a weapon system firing

Understanding why the subjects from the focus group struggle to see the link between their pure-tone thresholds and their AFFD is not difficult; there a very few face validity attributes of PTA as a measure of AFFD. An individual who does not have a background in audiological testing (referred to as the 'technically untrained observer' by Anastasi; 1988) may struggle to understand the link between listening to beeps in quiet and performance in an operational environment.

When reviewing available speech tests to decide which one should be further investigated as potential measure of AFFD the face validity of the test is one factor that should be considered. There are two main reasons for this. Firstly, a test which has very poor face validity is arguably less likely to be implemented. If it difficult to understand, at face value, why introducing a SIN would be better than the current test then it is possible that the 'technically untrained observed' will be less interested in pursuing its implementation. Secondly, by using a test that has high face validity, service personnel may begin to understand how their test performance links to their performance during operational scenarios. There are two potential benefits to this: 1) it may make it easier for the clinician to explain the test results and how they link to consequences in an operational environment; and 2) it may help personnel begin to understand and accept the impact their hearing loss is having on their job performance.

One of the key face validity aspects of a SIN test is the type of speech stimuli used. In order to assess this, a selection of SIN test stimuli representing a variety of stimuli types were rated by

military personnel in terms of their similarity to military communications. This pilot study is outlined in Section 5.2.4.

5.2.4 Pilot study: rating the similarity of speech test stimuli to command structure

In order to gather some information about the face validity of speech test stimuli a short survey was conducted which aimed to elicit their opinions about the similarities between speech test stimuli and military communication (Appendix C).

The survey included seven examples of SIN tests (BKB, CRM, HINT, MRT, Quick SIN, TDT and WIN, see Section 5.2.2 for information about these tests). These seven were selected to provide a general overview of the types of speech stimuli included in SIN tests. Participants were asked to rank, in their opinion, how relevant the speech test stimuli were in comparison to military communications. They were given three Likert scale ratings: 1) very relevant to military communication; 2) has some relevance to military communications; 3) no relevance to military communications.

Six personnel completed the survey. All the participants were infantry or combat support personnel, had experience in active service and had knowledge about command structure from their basic infantry training. Opportunistic sampling was used to recruit personnel; the author emailed all military personnel who had taken an advisory role at some point during the overall project and asked if they would be will to complete the survey. They were also asked if they could distribute the survey to two or three other infantry or combat support personnel.

The mean scores for each speech test were calculated and are shown in Table 5.3. The CRM, TDT and WIN speech stimuli showed the lowest average score. This indicates, albeit from a small sample and a survey which has not been validated, that these stimuli are thought to be very relevant to military communication format in terms of their face validity. Although the results of this survey cannot be used to draw any definitive conclusions about which speech test to choose they can be used to support any final decision in terms of whether a test is considered to hold high face validity to military communications.

Table 5.3 *Results of survey to investigate face validity of speech-in-noise tests for military AFFD testing*

Average score	**Speech test**						
	BKB	**HINT**	**QuickSIN**	**CRM**	**TDT**	**MRT**	**WIN**
1: very relevant to military communication ***2: some relevance to military communication*** ***3: no relevance to military communication***	2.3	1.8	2.0	1.2	1.3	2.0	1.2

5.2.5 Chosen speech-in-noise test: the coordinate response measure (CRM)

It has been decided that the CRM is a suitable test to investigate further as a measure of military AFFD

1. The results of the pilot study in Section 5.2.4 showed the CRM sentences structure "Ready call sign, go to colour number now" to have high face validity when compared to command structure. Although the CRM obtained the same score as the WIN the target words and carrier sentence used in the CRM is a much more similar format to military commands than the WIN stimuli.
2. The CRM speech material contains only words that military personnel will be familiar with and is therefore not testing vocabulary.
3. The sentences contain no syntactic or semantic information, meaning personnel's knowledge in this area will not impact their test results, minimising variation between individuals.
4. The speaker used for the CRM test will have an accent which is similar to Received Pronunciation to ensure the speech test is not measuring individual's ability to understand certain accents or dialects.

The test will initially be investigated in stationary speech-spectrum noise to minimise the number of auditory mechanisms involved during testing and therefore minimise the amount of potential for variation between subjects. Testing in this background noise also allows for a direct comparison to be made between the CRM and other speech tests which have been validated in stationary speech-spectrum noise during the validation stage of test development. The remainder of Chapter 5 goes through the stages for the development and evaluation of the CRM SIN test.

5.3 Study 2: Developing and recording CRM speech-in-noise test

5.3.1 Introduction

Section 5.2.5 has explained why the decision has been made to further investigate the CRM as a measure of AFFD for military personnel. Section 5.3, Study 2, will explain the process for recording the CRM, including the justification for creating a new recording of the test (Section 5.3.2), the format used to record the stimuli (Section 5.3.3) and the technical details of how the stimuli were recorded (Section 5.3.4).

5.3.2 Justification for re-recording the CRM

A version of the CRM has been recorded using a British male voice (Kitterick et al, 2010) but it was decided that by making some minor modifications to the target words the test could be made more relevant to British military communications. In addition, the use of externally sourced speech stimuli raised issues regarding the distribution of the speech test for use within the MoD. For these reasons it was decided that a new University of Southampton recording of the CRM would be created.

The call signs used in the Kitterick et al (2010) version of the CRM are those use in the United States Armed Forces and do not match the North Atlantic Treaty Organisation phonetic alphabet used by the British Army. Furthermore, disyllabic seven is included in the Kitterick et al (2010) recordings, making it possible for subjects to discriminate between words based on the number of syllables. For this reason it was decided to disregard seven from the new recording of the CRM and also to only include monosyllabic colours. Only disyllabic call signs were included, as these were in the majority (18 out of 26). The number zero is not included because the British Army pronounces this as ‘zero’, making it the only disyllabic number (British Army, 2011). In other speech tests it is pronounced ‘oh’ (e.g. the Triple Digit Test, TDT, Lutman et al, 2006) to avoid this problem but this would reduce the face validity of the test for military communications. The lists of stimuli included in the University of Southampton recording of the CRM are included in Table 5.4.

Table 5.4 *List of stimuli in the University of Southampton recording of the British English Coordinate Response Measure (CRM)*

Call signs		Colours	Numbers
Alpha	Oscar	Black	One
Bravo	Papa	Blue	Two
Charlie	Quebec	Brown	Three
Delta	Tango	Gold	Four
Echo	Victor	Green	Five
Foxtrot	Whiskey	Grey	Six
Hotel	X-Ray	Pink	Eight
Kilo	Yankee	Red	Nine
Lima	Zulu	White	Ten

5.3.3 Research objective 2

There is no 'knowledge gap' as such which relates to the research objective for Study 2.

Research objective 2: To design and record the British English version of the CRM.

5.3.4 Pilot study 1: deciding the format for recording sentences

An important technical detail that needed to be addressed prior to recording the CRM sentences was deciding on a format for recording the sentences. Previous recordings of the CRM had been done in full sentences resulting in numerous versions of the call signs, colours, numbers and carrier phrase throughout the test. For example, the call sign Alpha would have to be repeated 81 times in order record all the sentences containing each colour and number combination. This creates potential variation between sentences as there may be subtle differences in the intonation or vocal effort between sentence recordings. Recording the sentences in blocks of words may eliminate this variation but it was not known if this would cause disjointed sentences and distort the natural dynamics of speech.

A pilot study was run to investigate whether recording the sentences in blocks and then concatenating them to create any possible sentence would produce natural sounding speech or disjointed stimuli. To do this, a selection of the Kitterick et al (2010) CRM sentences were cut up into sections of 'Ready call sign', 'go to colour' and 'number now'. Eight example sentences were then concatenated by combining the separate target words of the sentence, with no two sentence-target words that were originally recorded together recombined. Seven individuals with English as a first language and assumed normal hearing (this was not tested for) were asked to listen to the eight concatenated sentences as well as the same eight sentences from the original

recording. They were asked to identify which sentence sounded like natural speech and which sounded like it had been modified (they were not told what the modification was). The results were averaged across the seven participants and showed that only 50% of the time (chance level) was the normal sentence was correctly identified, 37% of answers showed the participant mistaking the modified sentence for the natural sentence or vice versa and 13% of answers showed the participant could not tell the difference between the two sentences. Although this was a simple pilot experiment with a small sample it gave some confidence that splitting up the sentences and generating CRM sentences using components of the sentences does not result in noticeably different sounding sentences. The results of this small experiment resulted in the decision to record the CRM sentences as the separate components, thus minimising variation for each target word and reducing the volume of recording; if all possible sentences were recorded in their entirety this would result in 1458 sentences.

5.3.5 Recording the CRM stimuli

The CRM sentences were re-recorded in an anechoic chamber at the University of Southampton (see Figure 5.1). The anechoic chamber dimensions are 5.2 x 5.1 x 2.8 metres and the walls and ceiling are lined with open cell polyurethane foam wedges, 30cm long with a 30cm square base. The inner walls are isolated by a 25mm air gap from the outer structural wall. The room gives free-field conditions above approximately 250 Hz (Lower, 2014).

The speaker was a male with a standard southern English accent which was similar to Received Pronunciation (see Section 5.2.3 [2]). The speaker was aware of the intended use of the recordings and therefore understood the importance of maintaining a constant vocal effort throughout the recordings. The sentences were recorded in two formats. In the first format all the target words of the sentence were recorded separately, resulting in six recordings to make up one sentence. They were split into ‘Ready’, ‘call signs’, ‘go to’, ‘colours’, ‘numbers’ and ‘now’. The motivation for recording in this format was that it would then be possible to use the same recordings of ‘ready’, ‘go to’ and ‘now’ for every presented sentences, thus minimising variation between sentences. However, because it was not known if it would be possible to piece together sentences in this format and make them sound natural the sentences were also recorded as ‘Ready call sign’, ‘go to colour’ and ‘numbers now’, the same format which had been piloted, outlined in Section 5.3.3.

The speaker was asked to practice the intonation of the sentence and listened to several speakers saying the CRM sentences from the Kitterick et al (2010) recordings. They were then asked to practice saying the individual target words of the sentence whilst maintaining the same intonation

as if they were saying the whole sentence. They were instructed to maintain a natural intonation and loudness level, with the same vocal effort throughout the recordings. A minimum of six recordings of each sentence target word was taken to account for variations and allow for the best recording to be selected. The speaker was sat 0.75 metres from a Brüel and Kjær Precision Integrated Sound Level Meter, Type 2230, which was used as the recording device, due to its flat frequency response. This was connected a RME Babyface soundcard, plugged into a laptop running Microsoft Windows XP. The sounds were recorded as mono sound files at a 44100 Hz sampling rate using Adobe Audition. Figure 5.1 shows the recording set up.

Figure 5.1 *Equipment set up for recording CRM sentences*

5.3.6 Pilot study 2: evaluation and selection of CRM stimuli

To assess whether it would be possible to use the use the recording format, whereby the words 'Ready', 'call signs', 'go to', 'colours', 'numbers' and 'now' were recorded separately, a selection of these recordings were pieced together to form 10 example sentences. The same ten sentences were then generated using the second recording format of 'Ready call sign', 'go to colour' and 'number now'. Five native English speakers were asked to listen to the two versions of the sentences and to state which one sounded more like natural speech. These individuals had experience running speech test experiments and/or had previously participated in numerous experiments in which they listened to speech test stimuli and therefore had some understanding about speech test stimuli. The stimuli were sent over email and individuals listening to them using their own headphones/speakers. There was unanimous agreement that sentences made up of the separately recorded words sounded disjointed and unnatural. These recordings were discarded.

As stated in Section 5.3.4, a minimum of six recordings of each target word were made, allowing for the clearest version to be selected. For each target word the author made a decision about which of the six versions sounded the clearest and were most likely to sound natural when contained within a sentence. Once a single version had been chosen for each target word a set of 20 example sentences were concatenated. The set of sentences contained an example of all of the target words at least once. These sentences were then sent to the same native English speakers as in the above paragraph and they were asked to give their opinions in terms of the clarity of the recordings, the natural dynamics of the sentences and the pronunciation of the target words. It was agreed that the sentences sounded clear and natural but there were a few suggestions to change the pronunciation of some specific words (namely foxtrot, six and pink). These words were re-recorded using the same method outlined in Section 5.3.4 and used in the final version of the test.

5.3.7 Summary

Section 5.3 has outlined the process for recording the CRM test stimuli. At this stage a speech corpus containing eighteen dissyllabic call signs and nine monosyllabic colours and numbers has been created. The target words have been recorded with the preceding or proceeding carrier sentence, resulting in numerous versions of ‘Ready’, ‘go to’ and ‘now’, but this recording format resulted in the most natural sounding sentences. At this stage the intelligibility of the target words is unknown. Equal intelligibility within a speech corpus is an important factor when developing a SIN test; this is investigated in Section 5.4, Study 3.

5.4 Study 3: Equalising the intelligibility of the CRM in noise

5.4.1 Introduction

The next stage of developing a new SIN test is ensuring equal intelligibility across all stimuli; specifically for the CRM, the aim is to have minimal variation between the intelligibility of all the possible CRM sentences. Typically, SIN testing is performed using an adaptive procedure since this method allows for the rapid extraction of relevant information about an individual's SIN ability (Leek, 2001). The ultimate aim is to for the CRM test to be implemented in an adaptive procedure but prior to this it is necessary to check that it meets the key assumptions of SIN test stimuli.

According to Levitt (1971) there are four basic assumptions which are made when using an adaptive procedure. These are:

1. There must be a monotonic relationship between SNR and performance
2. Performance levels do not change during the course of the test (e.g. the test is not affected by experience or fatigue)
3. The PF has a specific parametric form (see Appendix B for further information about PFs)
4. Responses to each stimulus presentation are independent of the preceding and following stimuli

However, of these four assumptions Levitt (1971) states that only the first assumption is essential when carrying out an adaptive procedure. In order for a monotonic relationship to be observed it is necessary for the stimuli to be homogenous (Leek, 2001). Homogeneity refers to stability between performance levels across SNRs between target words. A monotonic relationship means that as the SNR increases so does performance level. These assumptions are important as it means that any sentence can be presented to the listener and the tester can be sure that the sentence is no easier or harder to understand because of the specific words it contains. The two assumptions are also closely related; if stimuli are heterogenic then this compromises monotonicity. If an 'easy' sentence is presented at a low SNR and achieves a correct response and a 'difficult' sentence is presented at a higher SNR and achieves an incorrect response this will result in a non-monotonic relationship caused by heterogenic stimuli. This leads to incorrect placement of trials in an adaptive procedure track, affecting the reliability of threshold estimations (Leek, 2001). Ultimately, the aim of Study 3 is to equalise the intelligibility of the CRM test material to ensure that when they are implemented in an adaptive procedure the measurement precision of the test is good. Measurement precision, in this context, refers to assessing the level of accuracy of the CRM adaptive procedure as a measure of SIN ability and is explained in more detail in Section 5.5.1 and Figure 5.13.

In order to assess the speech intelligibility of the CRM target words PFs must be measured for each of the stimuli. The target words refer to the individual call signs, colours and numbers, such as alpha, black and one. The target word groups refer to the groups of call signs, colours and numbers. A PF provides information about the relationship between a stimulus and the subjective responses. For more information about PFs please see Appendix B. For Study 3 the stimulus is the SNR at which the stimuli are presented (plotted on the x axis) and the subjective response is the percentage correct at any given SNR (plotted on the y axis) and a sigmoid curve ("S" shaped) is fitted to the data points.

To measure the intelligibility of the CRM target words the method of constant stimuli will be used. For this method the presentation level of a stimulus is not dependent on the preceding presentation level or response and it is therefore an appropriate method to use when the homogeneity of a speech corpus is unknown. This method also ensures that data is gathered for all points across the PF, from chance level to near 100% correct, which limits the amount of information that need to be interpolated to plot the PF, resulting in a better estimate of the shape of the PF.

The most important PF parameter for Study 3 is the location parameter of the PF for the individual target words. This is the only feature of the individual target words that can be amended to improve homogeneity. This is achieved by adjusting the amplitude of the stimuli which causes the PF to shift left (increasing the amplitude) or right (decreasing the amplitude), causing the location parameter to move accordingly. For this reason the main focus of the study is measuring and comparing the speech recognition 50% correct point (SRT 50) for individual target words. This will be read directly from the PF, rather than using the estimation of the location parameter of the PF ($\hat{\alpha}$), which is dependent on the upper and lower asymptotes and will therefore vary between target words groups. The guess rate for the call signs is lower (6%) than the colours and numbers (11%) because the call sign word group contains 18 target words compared to only 9 target words for the colours and numbers.

The slope of the target words is also an important feature; it is desirable for the PFs of target words to be steep, as this is a good indicator of the precision of the test score (Leensen et al, 2011a) and shallower slopes reduce the homogeneity of a speech corpus. When a slope is shallow a small change in the presentation level results in only a small change in the portion of correct responses. This increases the potential for threshold estimation error, since a range of presentation levels can elicit similar performance levels. For steeper curves, small changes in presentation level cause larger changes in performance level, reducing the threshold estimation error. Also, if target words have varying slopes then a change in SNR will elicit a different change

in performance levels between the target words, reducing the homogeneity of the stimuli. However, the slope of a target word is innate within the stimuli and therefore, unlike the threshold, is not altered through amplitude modifications. There is also no gold standard for defining how steep a slope must be for inclusion in a SIN test.

Study 3 will investigate the PFs of the CRM target words presented in stationary speech-spectrum noise, with the ultimate aim of ensuring the stimuli are homogenous, display monotonicity and have sufficiently steep slopes prior to implementation in an adaptive procedure. In the initial developmental stage of the CRM it was decided that a standardised noise which is not specific to any 'real world' listening environment should be used to provide a general measure of SIN ability. This type of noise involves limited auditory processing skills making it easier to interpret what is being measured during testing.

5.4.2 Research objective 3

Knowledge gap: No intelligibility measurements have been made for the University of Southampton CRM recordings presented in stationary speech-spectrum noise. It is not known if the target words meet the necessary assumptions for implementation in an adaptive procedure.

Research objective 3: To obtain speech intelligibility measurements for the individual call sign, colour and number target words of the CRM presented in stationary speech-spectrum noise and to adjust the stimuli amplitude to equalise the intelligibility of the CRM test material so the necessary assumptions for implementation in an adaptive procedure are met.

5.4.3 Method (route mean square equalised: sessions one and two)

A total of 20 normal hearing volunteers (10 male, 10 female, mean age 26 years) participated in the study, aged 18-35. Participants attended a total of three sessions. It was not possible to carry out a sample size calculation for this type of study; the aim was to gather enough data to obtain an accurate estimation of the PFs but there is no predefined sample size calculation method for this. The chosen sample size was selected in order to match or better previous studies which have successfully collected data from normal hearing listeners in order to estimate the PFs of speech stimuli (Ozimek et al, 2009; Smits et al, 2004). The method for all three sessions was identical but the amplitudes of the stimuli presented in session three had been adjusted based on the results of sessions one and two (see Section 5.4.5). Two participants withdrew from the experiment between sessions one and two.

All the participants were otologically normal. Normal hearing was defined as having hearing

thresholds of ≤ 20dB HL at 0.25, 0.5, 1, 2, 4 and 8 kHz. An otological health questionnaire was distributed to screen for tinnitus and ear disease, no potential participants were excluded as a result of this. Participation was on a purely voluntary basis, no payment was provided. Study 3 was approved by the University of Southampton (ERGO ref: 9762). The experiment was carried out in a sound proof booth and was run using a Mac laptop, running OS X Version 10.9.1. The stimuli were presented via an RME Babyface external sound card through Senheisser HDA 200 headphones. Calibration was performed in an artificial ear type 4153 using a flat plate coupler.

Using the method of constant stimuli, participants were presented with CRM sentences in stationary speech-spectrum noise (see Appendix D for details on how this was created) and were asked to respond to the call sign, colour and number they heard using a specially designed graphical user interface (Figure 5.2). If participants were not sure what they had heard they were asked to guess; they next sentence could not be played until a response had been given. Participants were given an opportunity to familiarise themselves with the interface and the stimuli by listener to between five and ten sentences at -1 dB SNR. In each session speech recognition for the call signs, colours and numbers was measured at seven SNRs (-1, -4, -6, -9, -12, -14 and -17); these were selected following a pilot study (n=6) conducted to estimate the SNRs required to elicit responses from chance level to near 100%. The order of the SNRs was randomly generated for each participant, using an online random number calculator (RandomOrg, 2014).

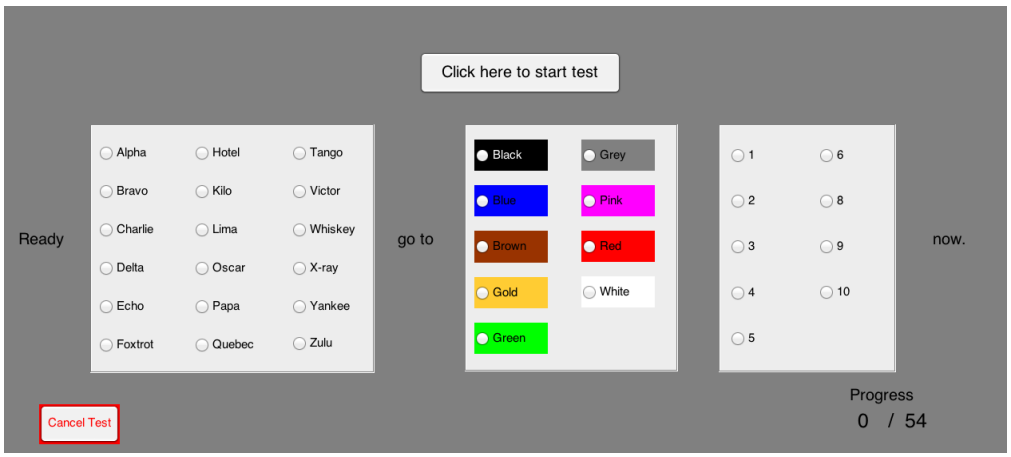


Figure 5.2 *A screen shot of the CRM graphical user interface*

The route mean square (RMS) amplitude of the CRM sentences were equalised (see Appendix E for details on how this was achieved). The CRM sentences were generated and scored using specifically designed MATLAB (R2013b) code (see Appendix F). At each SNR 54 sentences were presented; each call sign was presented three times, and each colour and number was presented nine times. For each presented sentence the target words were randomly selected. Figure 5.3 is a diagram showing the compilation process for each sentence. The 500ms of silence before and after each sentence allows sufficient time for the noise to be ramped to the maximum

presentation level prior to the speech starting to prevent forward masking. Moore (2008a) states that forward masking decays to 0 after at least 200ms, regardless of the initial amount of forward masking. A 15ms linear ramp was used. The 300ms of silence between “Ready call sign” and “go to colour” is to ensure the sentence sounds natural and not rushed. Appendix F contains the MATLAB (R2013b) code used to generate each sentence at the desired SNR.

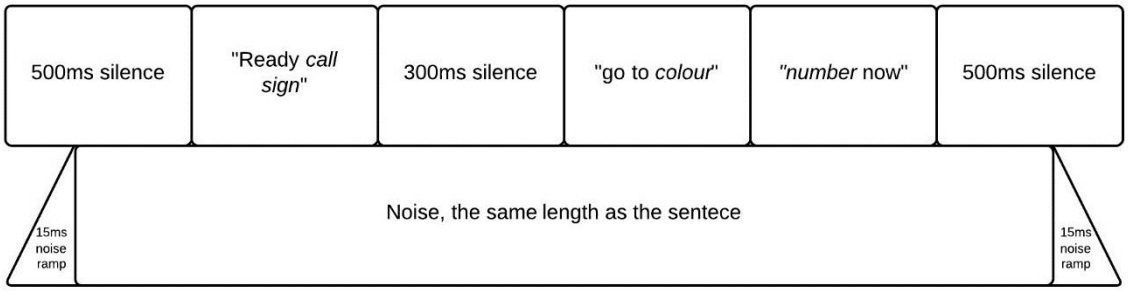


Figure 5.3 *The CRM sentence compilation*

The level of noise was kept at a constant value of 63dB A, thus the SNR was determined by varying the level of the CRM stimuli. For normal hearing listeners the presentation level does not have an effect on thresholds, providing the level is audible and comfortable (Smits et al, 2004). The chosen presentation level is based on that used by Ozimek et al (2009), in a similar study developing the Polish TDT.

The amount of gain required to be added or removed from the sentence was calculated using Equation 6. A custom noise file was created for every sentence presented; this ensures the noise and sentence are the same length and that the noise file is not identical for each sentence. This custom noise file is a randomly selected segment from within a 28 second long stationary speech-spectrum noise file (see Appendix D to see how this was generated).

Equation 6

$$required\ gain = \frac{\text{RMS noise}}{\text{RMS sentence}} \times 10^{\text{SNR}/20}$$

5.4.4 Results (sessions one and two)

The results were pooled across all the subjects, resulting in an averaged score for each of the CRM target words, at each SNR, for each session. The logistic functions model was used to fit a sigmoid curve to each CRM target words for sessions one and two, resulting in two PFs for each target word. The MATLAB Palamedes toolbox (Prins & Kingdom, 2009) was used to fit a logistic function to the data.

As explained in the background to PFs (Appendix B) there are a variety of different models used to fit a sigmoid curve to a PF. In order to check that a logistic function was a suitable model the goodness-of-fit for each PF for the session one data was checked. The goodness-of-fit shows how well the fitted logistic function accounts for the data and is given as a measure of deviance (pDev). According to Kingdom and Prins (2010, p.73), conventionally researchers agree that a fit is unacceptably poor if pDev is less than 0.05. For all the individual target words from session one the goodness-of-fit pDev was ≥0.61, with 29/36 target words obtaining a pDev score of ≥0.9. This consistently high pDev scores indicated that the logistic function fits the data well and was an appropriate function to use to analyse all subsequent data. Two examples are given in Figure 5.4 of the data points being fitted to the PF for the individual target word with the poorest goodness-of-fit (the number *six*, pDev=0.61) and an example of one of the best fits (the call sign *Victor*, pDev=1), giving some indication of how closely the raw data matches the fit. The SNRs tested for the number six did not span the whole range of values for the PF to be plotted, resulting in a poorer goodness-of-fit score. Although the data points still lie along the plotted line they are focused around the 100% correct point, not allowing for a full PF to be plotted.

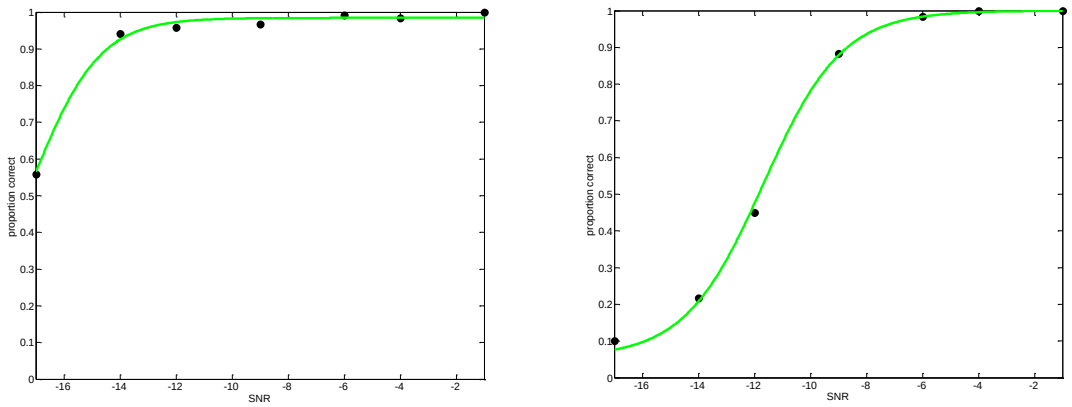


Figure 5.4 *Examples of a poor logistic function fit (left, data for target word 'six') and good logistic function fit (right, data for target word 'Victor'), averaged across participants from session one*

The percentage correct results at each SNR from sessions one and two were then averaged and PFs were plotted for each call sign, colour and number, shown in Figures 5.5, 5.6 and 5.7 respectively.

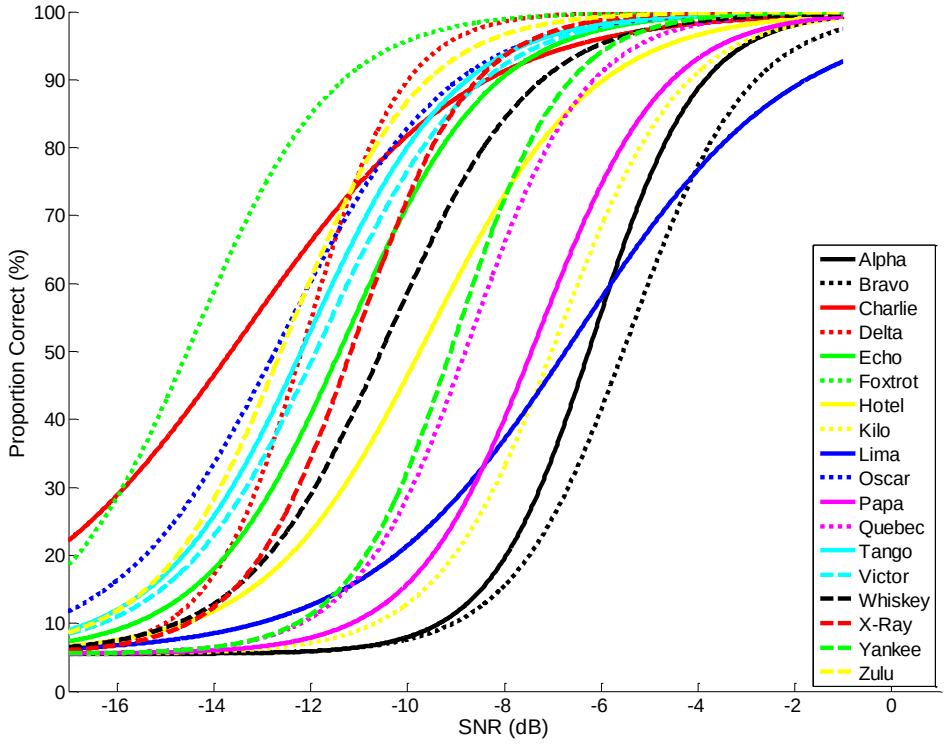


Figure 5.5 *Logistic functions for call sign target words (obtained from the mean scores at each SNR from sessions one, n=20, and two, n=18)*

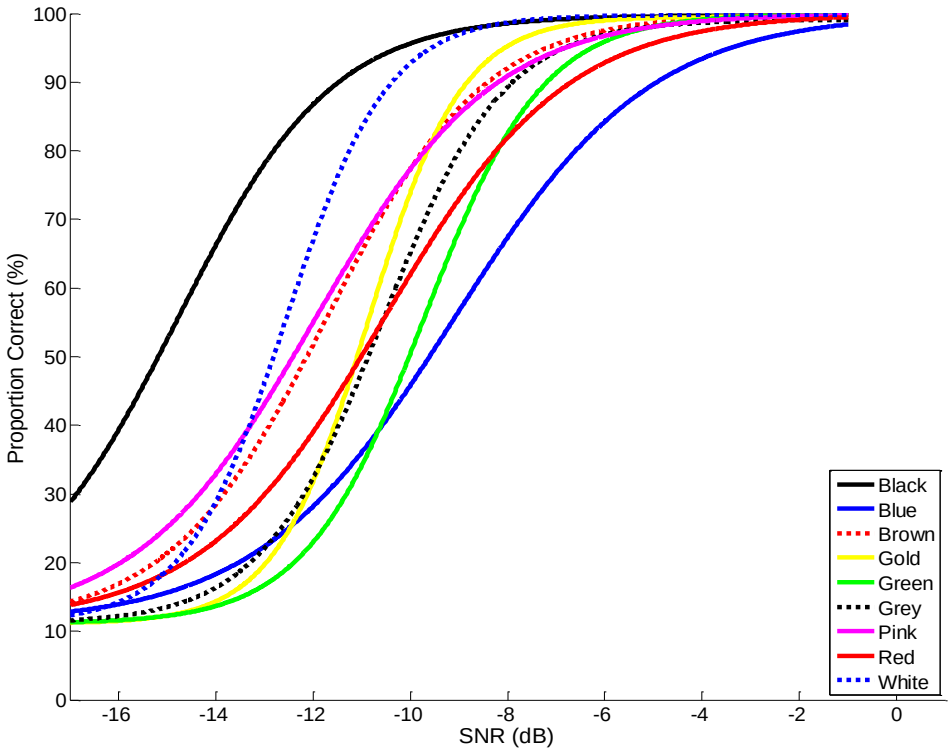


Figure 5.6 *Logistic functions for colour target words (obtained from the mean scores at each SNR from sessions one, n=20, and two, n=18)*

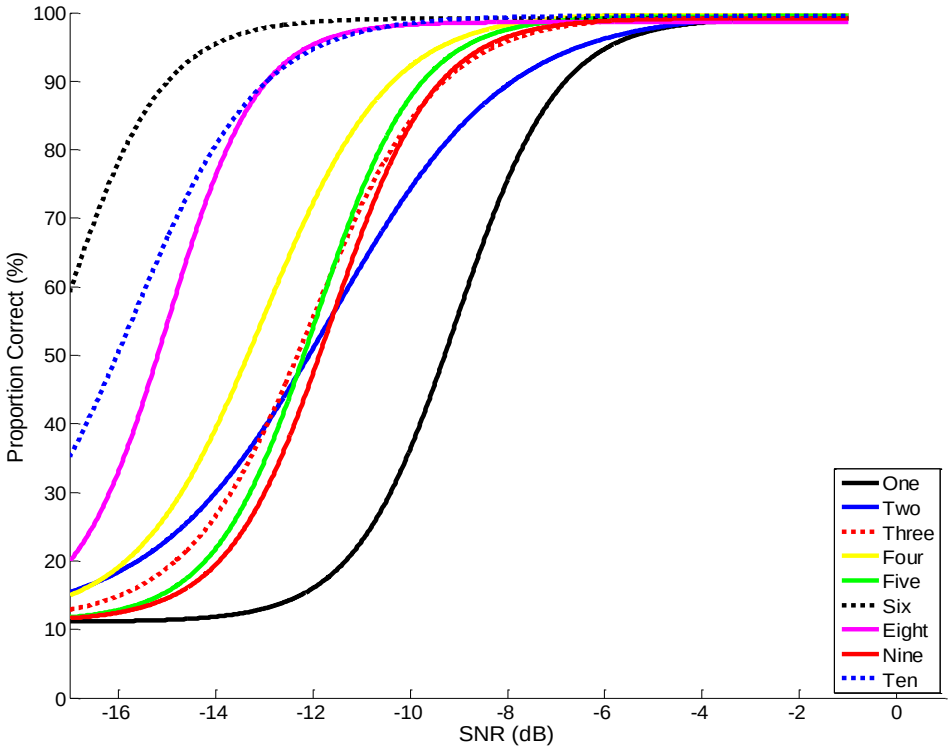


Figure 5.7 *Logistic functions for number target words (obtained from the mean scores at each SNR from sessions one, n=20, and two, n=18)*

The 50% correct speech recognition threshold (SRT 50) was read from the PF for each target word. These are shown for session one and session two in Table 5.5. In order to give a general idea about the magnitude of change in the SRT 50 scores between session one and two, two measures of repeat-reliability have been calculated. A more thorough introduction of repeat-reliability analysis is provided in Section 5.5.1. Two measures of repeat-reliability have been calculated, firstly the stability and secondly the repeatability correlation. The stability of the data was examined by looking at the magnitude of change in the SRT 50 score between the two sessions. The change in SRT 50 scores between repeats, averaged within target word groups, is very small (< 0.6dB) for all of the target word groups (call signs = 0.59dB, colours = 0.19dB, numbers = 0.41). Pearson's correlations have been calculated to show the level of agreement in the scores between the two sessions, for each of the target word groups. There was a strong (r = < .95) correlation between the two repeats for all the target word groups (call signs, r = .96; colours, r = .99; numbers r = .99).

The slope values are given in Table 5.5 for each of the target words. Visual inspection on the plotted PFs (Figures 5.5, 5.6 and 5.7) shows the majority of the slopes to be fairly steep. The mean slope values ($\hat{\beta}$) across session one and two are >0.7 for 28 of the 36 target words. The three target words displaying the shallowest slopes are Charlie, Lima and Blue, (< 0.5 $\hat{\beta}$). Prior to amplitude modification, which will only adjust the threshold, and not the slope, of the target

words, none of the target words were removed from the speech corpus based on their slope at this stage of test development. For this reason no further investigation was carried out regarding the slope. Following the results of session three the slope values will be further investigated to assess whether any target words have a slope too shallow for inclusion in an adaptive procedure.

There were large differences between the SRT 50 scores in the three groups. Overall participants found the call signs harder to hear than the colours and numbers. For this reason it was decided to try to equalise the SRT 50 scores within the target word groups, rather than across all the CRM target words. It would have been problematic to select an appropriate SRT 50 value which also matched all three groups since it would have meant large increases in amplitude for the call signs and large decreases in amplitude for the colours and numbers. This would have resulted in a large range of SNRs between target words for any one sentence, making it difficult to define the SNR of the presented sentence. For this reason the mean SRT 50 score for each target word group was used as the value to which individual target words were equalised. This results in the call signs, colours and numbers having different mean SRT 50 scores but does not compromise the goal of equalising the intelligibility of CRM sentences. By changing the amplitude of some of the individual target words that are easier or harder than others the amount of variation between thresholds is minimised. Section 5.4.5 explains the process for changing the stimuli amplitudes and the methods for re-measuring the PFs of the modified stimuli.

Table 5.5 *Thresholds (SRT 50) for the CRM target word (highlighted red= amplitude altered before session three)*

Call signs	Mean threshold session 1& 2 (SRT 50) = -10.26 dB								
	Alpha	**Bravo**	**Charlie**	**Delta**	**Echo**	**Foxtrot**	**Hotel**	**Kilo**	**Lima**
Session 1 SRT 50 (dB)	-6.19	-5.24	-12.81	-12.29	-11.21	-14.26	-10.01	-6.81	-6.17
Session 2 SRT 50 (dB)	-6.33	-6.00	-14.43	-12.11	-11.60	-14.81	-9.47	-9.40	-7.35
Difference in SRT 50 between Session 1 &2 (dB)	0.14	0.75	1.62	0.18	0.39	0.55	0.54	2.59	1.18
Mean SRT 50 session 1 & 2 (dB)	-6.26	-5.62	-13.62	-12.20	-11.40	-14.54	-9.74	-8.11	-6.76
Difference in SRT 50 from group mean (dB)	-4.00	-4.64	3.36	1.94	1.14	4.28	-0.52	-2.15	-3.50
Session 1 ($\hat{\beta}$)	0.91	1.11	0.43	1.01	0.65	0.74	0.58	1.04	0.51
Session 2 ($\hat{\beta}$)	0.97	0.76	0.45	1.03	0.72	0.71	0.61	0.35	0.41
Call signs cont.	**Oscar**	**Papa**	**Quebec**	**Tango**	**Victor**	**Whiskey**	**X-Ray**	**Yankee**	**Zulu**
Session 1 SRT 50 (dB)	-12.57	-7.37	-8.88	-11.34	-11.85	-10.34	-10.74	-8.79	-12.51
Session 2 SRT 50 (dB)	-12.90	-7.52	-8.65	-12.99	-11.94	-10.72	-11.59	-9.43	-12.70
Difference in SRT 50 between Session 1 &2 (dB)	0.33	0.16	0.23	1.65	0.09	0.38	0.84	0.65	0.18
Mean SRT 50 session 1 & 2 (dB)	-12.73	-7.45	-8.77	-12.16	-11.89	-10.53	-11.17	-9.11	-12.60
Difference in SRT 50 from group mean (dB)	2.48	-2.81	-1.49	1.90	1.63	0.27	0.91	-1.15	2.34
Session 1 ($\hat{\beta}$)	0.57	0.91	0.92	0.72	0.71	0.76	0.82	0.93	0.81
Session 2 ($\hat{\beta}$)	0.62	0.67	0.78	0.75	0.60	0.65	0.96	0.90	0.69
Colours	Mean threshold session 1& 2 (SRT 50) = -11.67 dB								
	Black	**Blue**	**Brown**	**Gold**	**Green**	**Grey**	**Pink**	**Red**	**White**
Session 1 SRT 50 (dB)	-14.80	-9.37	-11.81	-10.99	-9.97	-10.58	-12.14	-11.16	-12.54
Session 2 SRT 50 (dB)	-15.50	-9.83	-12.43	-11.21	-10.07	-11.10	-12.67	-10.87	-13.07
Difference in SRT 50 between Session 1 &2 (dB)	0.35	0.23	0.31	0.11	0.05	0.26	0.26	0.14	0.27
Mean SRT 50 session 1 & 2 (dB)	-15.15	-9.60	-12.12	-11.10	-10.02	-10.84	-12.41	-11.02	-12.80
Difference in SRT 50 from group mean (dB)	3.48	-2.08	0.45	-0.57	-1.65	-0.83	0.73	-0.66	1.13
Session 1 ($\hat{\beta}$)	0.58	0.48	0.67	0.97	0.94	0.72	0.55	0.53	0.94
Session 2 ($\hat{\beta}$)	0.74	0.51	0.61	1.22	0.76	0.87	0.56	0.56	1.01
Numbers	Mean threshold session 1& 2 (SRT 50) = -13.31 dB								
	One	**Two**	**Three**	**Four**	**Five**	**Six**	**Eight**	**Nine**	**Ten**
Session 1 SRT 50 (dB)	-9.00	-11.69	-12.32	-12.92	-12.02	-17.35	-14.84	-11.86	-15.94
Session 2 SRT 50 (dB)	-9.81	-12.44	-12.31	-13.70	-12.34	-17.44	-15.58	-11.93	-16.11
Difference in SRT 50 between session 1 &2 (dB)	0.81	0.75	0.01	0.79	0.33	0.09	0.74	0.07	0.17
Mean SRT 50 session 1 & 2 (dB)	-9.41	-12.06	-12.31	-13.31	-12.18	-17.39	-15.21	-11.90	-16.03
Difference in SRT 50 from group mean (dB)	-3.90	-1.25	-1.00	0.00	-1.13	4.08	1.90	-1.41	2.72
Session 1 ($\hat{\beta}$)	0.64	0.52	0.91	0.79	0.91	0.85	1.19	1.03	0.66
Session 2 ($\hat{\beta}$)	0.70	0.61	0.68	0.80	1.04	1.23	1.03	0.98	0.89

5.4.5 Method (amplitude equalised: session three)

Any target word with an SRT 50 score that differed from the target word group mean by ±1.5dB had its amplitude increased or decreased by the difference, to the nearest half decibel (shown in Table 5.5, highlighted red). This the same value used by Ozimek et al (2009) in the development of the Polish TDT for evaluating the intelligibility functions of triplet lists. This criterion resulted in no significant differences between the speech intelligibility of the triplet lists when implemented in an adaptive procedure. The amplitude was changed in Adobe Audition and the modified stimuli were saved as new wav files and used in session three.

After the stimuli had been modified based on the SRT 50 scores, the exact same method used in sessions one and two was carried out, as outlined in Section 5.4.3, on the same sample used in session two (n=18). The hearing test was not repeated; when questioned, no participants felt their hearing had changed since their previous session. All participants repeated the third session within two months of their first and second sessions.

5.4.6 Results (session three)

Results were analysed using the same methods used in Section 5.4.4. Logistic functions were plotted for each of the individual CRM target words (shown in Figures 5.8, 5.9 and 5.10), using the data pooled across all the participants. The goodness-of-fit was calculated for each of the CRM target words to check the logistic function was still a suitable model to fit to the data. For all of the individual target words from session three the goodness of fit pDev was $\geq$0.63, with 30/36 target words obtaining a pDev score of >0.9, providing evidence that the logistic function is accounting for the data and is an appropriate fit. For each CRM target word group the SRT 50 was calculated and is reported in Table 5.6. To help the reader see the change in shape of the PF between sessions one and two compared to session three, the PFs are plotted adjacently in Appendix G.

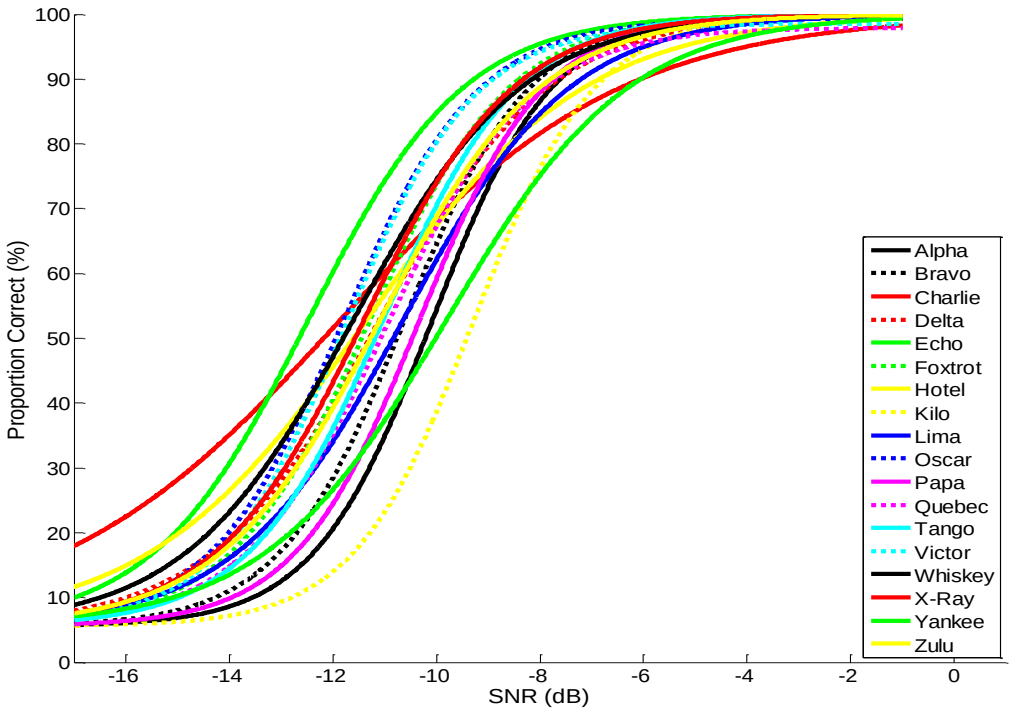


Figure 5.8 *Logistic functions for call sign target words (obtained from the mean scores at each SNR from session three)*

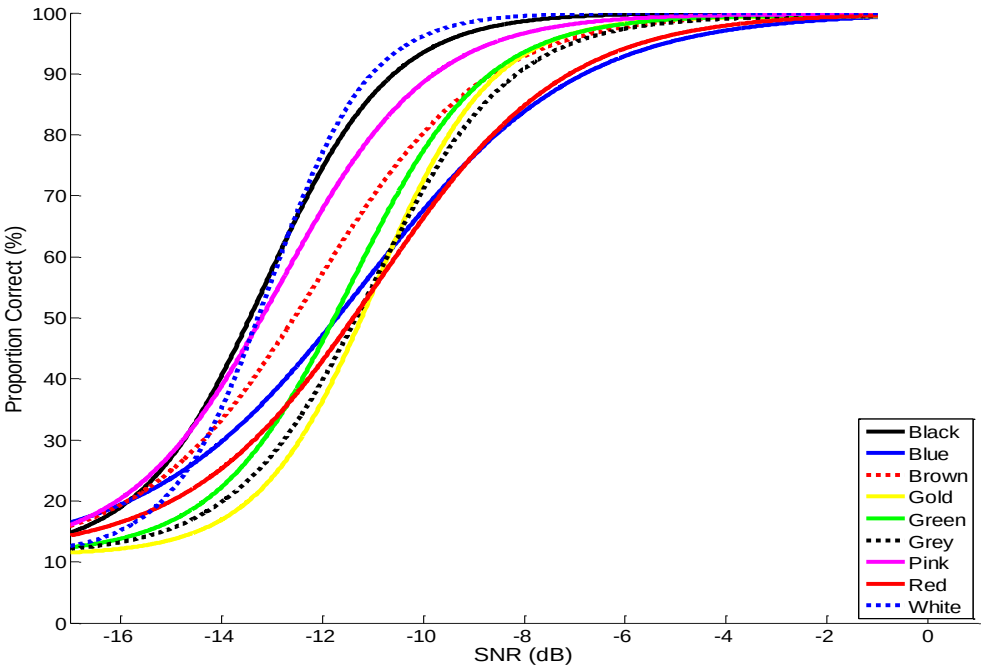


Figure 5.9 *Logistic functions for colour target words (obtained from the mean scores at each SNR from session three)*

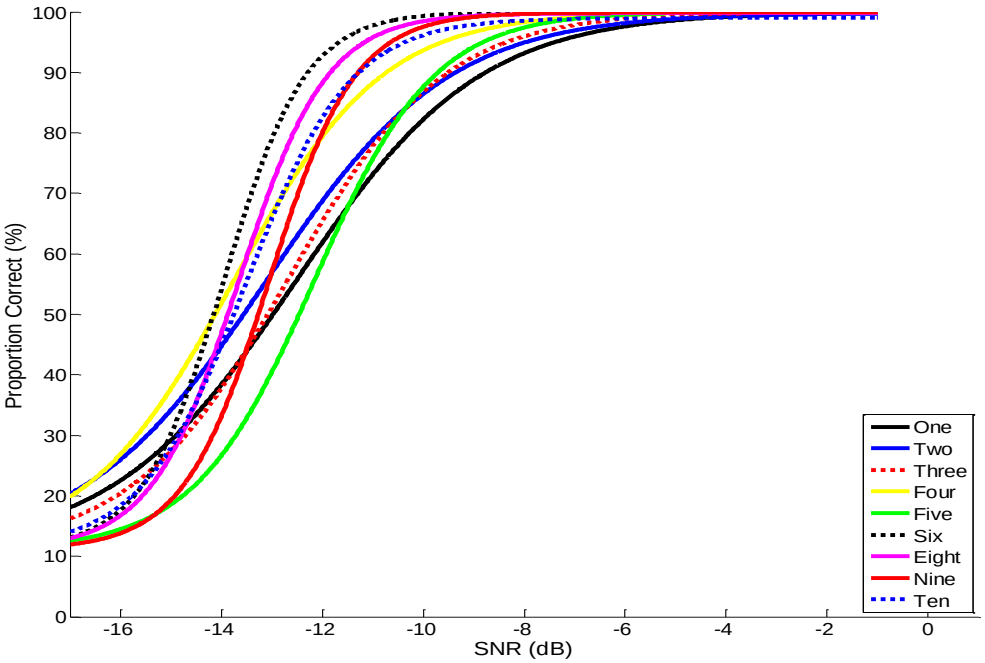


Figure 5.10 *Logistic functions for number target words (obtained from the mean scores at each SNR from session three)*

As explained in Appendix B (background to PFs), the values given for the threshold and slope of the PF are estimates of the "true" value and it is not possible to know the exact value, since this is a theoretical number. The method of bootstrap analysis allows for estimation of the error of the observed value and calculates the 95% confidence intervals for the threshold and slope of the logistic functions. For this method a random set of hypothetical data, based on the recorded experimental data is generated. For each set of hypothetical data, a logistic function is fitted and estimates of the threshold and slope are generated. Using the simulated sets (400 in total), the estimated 95% confidence interval of the true thresholds is calculated for each of the individual target words, reported in Table 5.6. Due to the unstable nature of the slope of the PFs, bootstrap analysis does not generate meaningful data regarding the error in the slope. As an alternative approach, the recorded slope values for sessions one, two and three of Study 2 can be compared. There is only a small amount of variation in the slope values between individual target words between the three sessions. The mean difference (± one standard deviation) between minimum and maximum slope values, across the three sessions, for each target word group is: callsigns=0.2 (±0.1), colours=0.3 (±0.3) and numbers=0.21 (±0.1).

As explained in the introduction, the steepness of a target word intelligibility function affects the measurement precision; a steeper slope generates more precise threshold estimation. A plot of the estimated slope value and the 95% confidence intervals of the thresholds clearly demonstrate this relationship. There is no definitive slope 'cut off value' for speech stimuli to be included in SIN test but Figure 5.11 allows for measurement precision and slope to be assessed together to

decide what is an acceptable value of error and the corresponding slope value. Figure 5.11 indicates that slope values of roughly 0.5 and above have a measurement error of around ± 3dB or below, which includes the majority of target words and highlights *Charlie* and *Hotel* as outliers, with slopes below 0.5 and a 95% confidence interval of threshold estimation above 3.5dB.

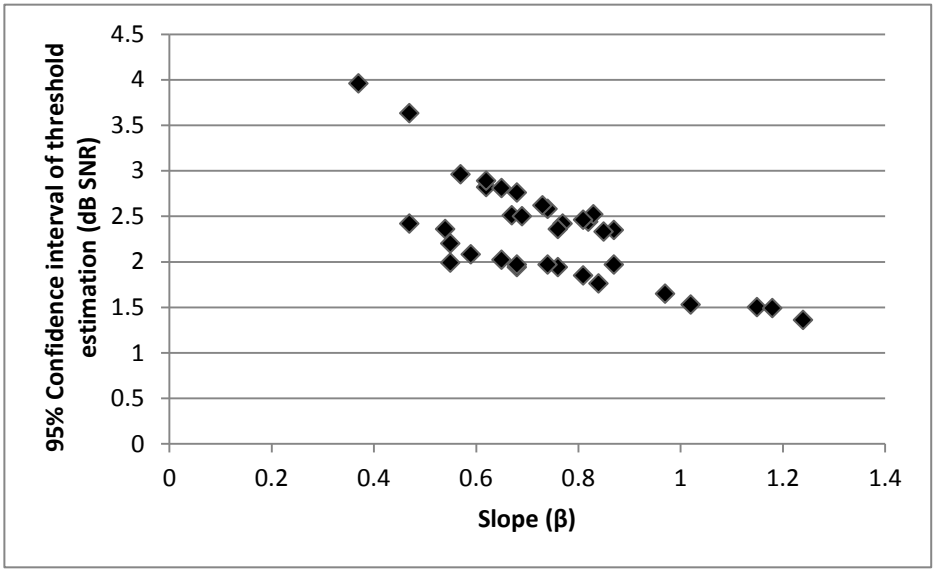


Figure 5.11 *Graph showing the relationship between the 95% confidence interval of threshold estimation and the slope value of each target word*

Table 5.6 *Threshold (SRT 50) scores, threshold 95% confidence intervals and slope for each CRM target word*

Call signs	**Alpha**	**Bravo**	**Charlie**	**Delta**	**Echo**	**Foxtrot**	**Hotel**	**Kilo**	**Lima**
Session 3 SRT 50 (dB)	-10.21	-10.76	-12.18	-11.29	-12.65	-11.44	-11.60	-9.41	-10.84
Threshold 95% confidence interval	2.35	2.44	3.96	2.82	2.51	2.58	3.63	2.33	2.76
Slope ($\widehat{\beta}$)	0.87	0.82	0.37	0.62	0.67	0.74	0.47	0.85	0.68
Call signs cont.	**Oscar**	**Papa**	**Quebec**	**Tango**	**Victor**	**Whiskey**	**X-Ray**	**Yankee**	**Zulu**
Session 3 SRT 50 (dB)	-11.94	-10.47	-11.13	-11.19	-11.86	-11.80	-11.57	-10.01	-11.27
Threshold 95% confidence interval	2.42	2.52	2.62	2.36	2.46	2.89	2.50	2.96	2.81
Slope	0.77	0.83	0.73	0.76	0.81	0.62	0.69	0.57	0.65
Colours	**Black**	**Blue**	**Brown**	**Gold**	**Green**	**Grey**	**Pink**	**Red**	**White**
Session 3 SRT 50 (dB)	-13.44	-11.72	-12.56	-11.23	-11.77	-11.33	-13.21	-11.39	-13.28
Threshold 95% confidence interval	1.85	2.42	2.08	1.97	1.94	1.97	1.94	2.36	1.53
Slope ($\widehat{\beta}$)	0.81	0.47	0.59	0.87	0.76	0.74	0.68	0.54	1.02
Numbers	**One**	**Two**	**Three**	**Four**	**Five**	**Six**	**Eight**	**Nine**	**Ten**
Session 3 SRT 50 (dB)	-12.98	-13.56	-13.08	-14.12	-12.46	-14.16	-13.87	-13.25	-13.75
Threshold 95% confidence interval	2.20	1.99	2.02	1.97	1.76	1.36	1.50	1.49	1.65
Slope ($\widehat{\beta}$)	0.55	0.55	0.65	0.68	0.84	1.24	1.15	1.18	0.97

As a result of equalising the intelligibility with the target word groups the standard deviation of the SRT 50 scores within each target word group decreased between sessions one and two and session three, shown in Table 5.7. All the target words SRT 50 score (± 95% CI, Table 5.6) are within ±1.5 the corresponding target word group mean SRT 50.

Table 5.7 *Target word group mean SRT 50 score details for sessions 1 & 2 (combined) and session 3*

		Call signs	***Colours***	***Numbers***
Session 3	**Target word group mean SRT 50 (dB)**	-11.20	-12.21	-13.47
	Target word with lowest SRT 50 (*'easiest',* dB*)*	-12.65 *(Echo)*	-13.44 *(Black)*	-14.16 *(Six)*
	Target word with highest SRT 50 (*'hardest',* dB*)*	-9.41 *(Kilo)*	-11.23 *(Green)*	-12.46 *(Five)*
Sessions 1 & 2 target word group SRT 50 one standard deviation		2.65	1.68	2.47
Session 3 target word group SRT 50 one standard deviation		0.81	0.91	0.57

Finally, confusion matrices were generated for each of the target word groups, using results pooled from all three sessions and for each SNR tested (Figure 5.12). These show participants responses (vertical) compared to the presented stimuli (horizontal), as a percentage of all incorrect responses across all presentations (calculated by the total presentations of each call sign (1176), colour and number (2352) minus the number of correct responses for the target word, multiplied by 100). A grey scale has been applied to each column of the confusion matrices (excluding the main diagonal) whereby any percentages between 0-10% are light grey, any between 11-20% are dark grey and any over 20% are black, allowing the reader to see at a glance which target words are being confused more commonly than others.

A series of chi squared tests of goodness-of-fit were performed to determine whether the observed incorrect responses for each target word were evenly distributed. Incorrect responses were not equally distributed for any of the target words (significant at p=<0.001 across all target words). This suggests that when a participant does not respond correctly they are not selecting an alternative answer entirely at random. There are two possible reasons for this. The first is simply that participants always selected the first answers in the target word group or the answers at the top of the screen, when they were unsure, resulting in a higher number of incorrect responses for *Alpha, Hotel, Tango, Black, Grey, One* and *Six*. However, the confusion matrices do not reveal these target word as being consistently confused more than others so this cannot explain the abnormal distribution of errors. The second reason for confusion is that words which are similar phonetic content are being confused. There does appear to be some patterns of phonetic confusion such as *Hotel/Victor* (presented/responded) share the phoneme /t/(25%),

Whiskey/Oscar share /k/(42%), *Black/Brown* share /b/(35%) and *Ten/Two* share /t/(23%). It is difficult to decide at what point consistent phonetic confusion becomes problematic and impacts test performance. The two words which are being confused more often than any others are *Whiskey* and *Oscar.* When *Whiskey* is presented, 42% of the incorrect responses are for *Oscar*. Interestingly, reverse confusion is not equally high, when *Oscar* is presented only 6% of incorrect responses are for *Whiskey*. At this stage it is not known whether these confusions will affect the measurement precision of the CRM adaptive procedure and for this reason no words will be excluded from the CRM test based on the confusion matrix results.

Participant Response \ Presented Stimuli	A	B	C	D	E	F	H	K	L	O	P	Q	T	V	W	X	Y	Z
A		2	2	1	4	2	1	2	3	3	4	3	1	1	1	1	2	2
B	3		4	4	3	3	1	5	5	2	14	4	4	3	2	3	2	2
C	3	3		5	6	13	2	5	3	11	5	4	7	3	3	10	4	5
D	9	5	8		8	5	21	4	6	6	5	6	16	21	5	5	7	7
E	12	15	5	3		4	2	10	7	3	8	7	5	3	2	4	27	3
F	4	2	10	7	3		2	3	4	11	3	2	4	4	4	20	2	7
H	5	5	4	11	4	6		5	3	5	4	6	11	17	3	4	3	4
K	5	10	6	3	6	7	2		10	4	8	8	5	3	2	5	3	7
L	9	11	6	4	10	8	2	14		5	11	10	6	6	5	7	8	7
O	7	4	6	10	4	6	7	6	7		7	6	6	6	42	12	3	14
P	7	9	4	4	9	4	2	6	8	4		8	6	5	4	5	5	4
Q	3	4	1	1	4	4	2	4	6	1	3		2	1	2	4	4	2
T	6	8	14	15	11	5	18	9	5	7	7	8		13	7	6	10	7
V	8	8	10	17	8	12	25	7	7	6	7	6	12		4	5	7	9
W	3	5	6	6	7	7	6	5	7	9	6	6	4	6		3	5	9
X	5	3	4	5	4	4	3	3	5	12	3	7	5	5	6		6	7
Y	7	3	6	3	6	5	2	4	6	3	3	6	6	3	2	4		6
Z	3	3	3	2	1	1	1	7	8	8	3	2	3	2	6	1	2	

Participant Response \ Presented Stimuli	Black	Blue	Brown	Gold	Green	Grey	Pink	Red	White
Black		6	23	8	6	9	7	8	12
Blue	13		17	19	24	11	13	20	14
Brown	35	14		18	17	19	18	18	18
Gold	10	21	10		11	7	9	10	16
Green	6	19	5	9		11	13	13	6
Grey	6	8	12	7	8		12	12	6
Pink	10	8	9	10	8	7		10	10
Red	11	15	17	13	18	28	18		19
White	7	9	7	16	7	8	8	8	

Participant Response \ Presented Stimuli	One	Two	Three	Four	Five	Six	Eight	Nine	Ten
One		10	15	16	18	13	13	16	12
Two	15		25	18	14	17	20	15	23
Three	21	25		29	17	28	24	23	17
Four	21	15	20		12	12	15	15	13
Five	9	5	8	6		8	4	11	6
Six	3	4	2	5	3		6	3	4
Eight	8	9	13	12	10	8		12	17
Nine	19	8	12	9	23	5	10		8
Ten	4	24	4	5	3	8	8	5	

Figure 5.12 *Confusion matrices for the CRM target words across all three sessions (top: call signs, letters refer to first letter of call signs, bottom left: colours, bottom right: numbers). Each column totals 100%, calculated as a percentage of all incorrect responses across all presentations across all Study 3 sessions.*

5.4.7 Discussion

The aim of Study 3 was to ensure the CRM stimuli met the necessary assumptions for inclusion in an adaptive procedure. The only essential assumption when using an adaptive procedure is that there is a monotonic relationship between the SNR and performance (Levitt, 1971). In order for a monotonic relationship to be observed it is necessary for the stimuli to be homogenous (Leek, 2001). The study measured the intelligibility functions of the CRM target words in stationary speech-spectrum noise and modified the stimuli amplitude to achieve homogeneity within target word groups.

Homogeneity of speech stimuli is the first key assumption for speech material used in an adaptive procedure. Following the amplitude modifications of the target words between sessions one and two and session three the standard deviation of the SRT 50 scores within the target word groups reduced. Following session three the SRT 50 score of any target word is now within ±1.5dB (± 95% CI) of the corresponding target word group. Prior to implementation in an adaptive procedure it is difficult to comment on whether the chosen inclusion criterion of ±1.5dB (± the target words SRT 50 95% CI) is sufficient for the target words to be considered homogenous. If measurement precision values are very poor for the adaptive procedure method, some of the target words that deviate the most from their group mean or have the largest 95% confidence intervals may need to be considered for removal from the speech corpus. Since no single target word deviated from this criterion, no target words have been excluded at this stage and homogeneity is being assumed.

Monotonicity of the speech stimuli is the second key assumption when designing an adaptive procedure SIN test. It is clear from the PFs in Figures 5.8, 5.9 and 5.10 that the target words independently behave in a monotonic manner, an increase in SNR causes an increase in proportion correct and vice versa. In this particular experiment the monotonicity of the speech corpus when scored as a whole sentence has not been assessed. However, since the individual target words are homogenous a monotonic relationship can be assumed. If all possible CRM sentences are homogenous then an increase or decrease in SNR will result in a corresponding change in proportion of correct responses.

The steepness of the PF slopes is the final factor to consider prior to adaptive procedure implementation. As shown in Figure 5.11 target words with shallow slopes display poorer measurement precision than those with steep slopes. As with the homogeneity of the stimuli there is no set value for how steep the slope of a target word to be for it to be suitable for inclusion in an adaptive procedure. When assessing the relationship between slope and measurement precision the target words which showed the shallowest slope and largest SRT 50 95% confidence intervals were *Charlie* (slope ($\hat{\beta}$) = 0.37, 95% CI=3.96dB) and *Hotel* (slope ($\hat{\beta}$)= 0.47,

95% CI=3.63dB). *Blue* also displayed a shallow slope ($\hat{\beta}$ 0.47) but a 95% confidence interval similar to the other target words (2.42dB). The remaining target words displayed slopes > 0.5 and 95% confidence intervals of < 3dB. At this stage it is not possible to know if the shallow slope target words *Hotel*, *Charlie* and *Blue* will have a significant effect on the measure precision of the CRM adaptive procedure test. No target words have been removed based on their slope values; if the measurement precision of the adaptive procedure is shown to be poor one consideration would be the removal of the words with the shallowest slopes.

When designing a SIN test one factor to consider is the scoring method used. For a response to be considered 'correct' some tests require the listener to correctly hear the entire sentence (such as the TDT) and others are only required to correctly hear key words (such as the BKB sentences). The most common scoring method used for the CRM test is only scoring responses to the colour and number aspect of the sentence. In the past the CRM test has mainly been used to test speech masked by speech and the call sign has been used to indicate to the listener which target phrase to lock onto, rather than as an additional word for identification (Eddins & Liu, 2012; Bolia et al, 2000; Brungart, 2001a; Brungart, 2001b). It has also been used for speech localisation experiments, again using the call sign as the target phrase for the listener to attend to (for example Rothpletz et al, 2011). There does not appear to be any previous work using the CRM as a screening measure. The ultimate aim of this project is to introduce the CRM SIN test as a measure of AFFD, not as a localisation, attention or masking effect test, as has been the case for previous usage. As a measure of AFFD the CRM needs to hold high face validity to command structure (see Section 5.2.3 [6] and Section 5.2.4), whereby all key word in the command need to be correctly identified. Scoring the CRM on correct identification of all three target words would arguably hold higher face validity compared to omitting the call sign. In a real world combat scenario if personnel did not know who a command was being issued to this would impact communication success. However, there is concern that the introduction of the call sign target word group may reduce measurement precision of the adaptive procedure test. Firstly, the call sign target word group displays the largest range of SRT 50 scores, making it the least homogenous, which may impact on the monotonicity of the speech corpus. Secondly, the target words with the shallowest slopes and highest measurement error (*Charlie* and *Hotel*) are within the call sign target word group, which may impact the overall precision of the test. Rather than make any assumptions at this stage about the suitability of the call sign target word group for inclusion in an adaptive procedure test, two scoring methods will be further investigated, the correct identification of: 1) all three target words and 2) only the colour and number.

5.4.8 Conclusion

The CRM sentence materials are ready for implementation in an adaptive procedure. Homogeneity of stimuli and a monotonic relationship between presentation level and response has been investigated and it is concluded that all the CRM target words sufficiently satisfy these criteria to be used in the CRM adaptive procedure test. It is currently not known if certain target words which display shallow slopes and higher levels of measurement error will affect the measurement precision of the CRM adaptive procedure test. If the test shows poor measurement precision the inclusion of these target words will need to be reconsidered. It is also not known if a scoring method which incorporates the call sign target word group will reduce the measurement precision of CRM test. For this reason the measurement precision of two scoring methods (correctly identifying all three target words or only the colours and numbers) will be compared in Study 4.

5.5 Study 4: Exploring the measurement precision of the CRM implemented in an adaptive procedure

5.5.1 Introduction

Study 3 has concluded that CRM sentence stimuli are now ready for implementation in an adaptive procedure. Adaptive procedures allow for the rapid extraction of relevant measurements from a PF (Leek, 2001). A SIN test needs to be able to accurately and quickly gather information about an individual's SRT. According to Leek (2001) adaptive procedures allows for relevant observations about the PF to be made with maximum efficiency but without sacrificing accuracy and the method is commonly used for measuring SIN thresholds.

The final two stages of developing the CRM as a measure of AFFD involve assessing the measurement precision and the predictive validity of the CRM adaptive procedure test. Study 4 will focus on exploring the measurement precision of the test and steps towards exploring the predictive validity of the CRM as a tool for measuring AFFD is addressed in Chapter 6. Measurement precision, in this context, refers to assessing the level of accuracy of the CRM as a measure of SIN ability and is determined by two key pieces of information: 1) reliability and 2) the validity. Reliability and validity encompass a number of aspects which are mapped out in Figure 5.13. The terminology used in Figure 5.13 is based on that used by Summerfield et al (1994) in their paper looking at test-retest reliability of the IHR McCormick Automated Toy Test. The purpose of Study 4 is to investigate the aspects of measurement precision shown in Figure 5.13 for the CRM with two different scoring methods, in order to determine whether it is appropriate to further explore the use of the CRM adaptive procedure test (in one of the test condition formats) as a measure of AFFD. If the CRM test does not display good measurement precision then it will not be worth exploring its use any further.

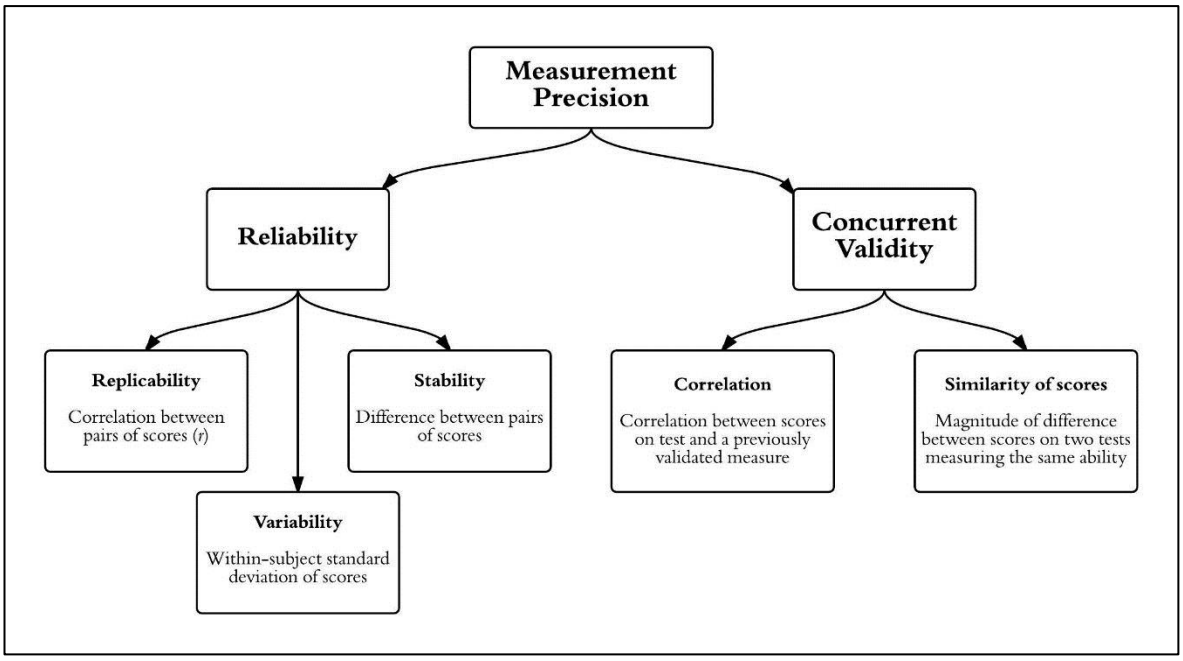


Figure 5.13 *Mind map showing the relationship between different aspects of measurement precision*

The first aspect of measurement precision is the **reliability** of a test. One characteristic of a 'perfect test' is a measurement tool that provides identical scores when the same subject is tested under the same conditions on two different occasions (Summerfield et al, 1994). However, this 'perfect test' does not exist; in reality there is always some degree of variation between repeated measures. There are two types of uncertainty in any measurement which cause this variation between repeats: 1) random and 2) systematic (Bell, 2001). Random error occurs when repeated measurements produce randomly different results each time. The amount of random error indicates how accurate a measurement tool is. Systematic error occurs when repeated measurements are affected by the same influence each time. The amount of systematic error is influenced by factors such as learning effect or fatigue (Bell, 2001).

Investigating the reliability of a test involves evaluating how scores change between repeats. The following three paragraphs describe the three measures of test reliability (replicability, variability and stability, as shown in Figure 5.13), explaining how they can be used to assess test reliability and how they indicate the amount of random and systematic variation within a test.

Replicability of a test is measured by the correlation between repeated scores. A high correlation between scores means that one score can be predicted accurately from the other. Correlation assesses the change in rank order of subjects between repeats. However, the correlation value between repeats has two major limitations for use as a measure of test reliability. Firstly, it is possible to obtain a high correlation value without the test scores being the same; this is caused by systematic error. If magnitude of the change between repeats is similar across the sample

(possibly caused by a learning effect or fatigue) then the rank order of the subjects does not change between repeats, resulting in a high correlation, despite differences in the scores. Secondly, a low correlation value does not necessarily indicate poor replicability. Correlation calculations strongly depend on the test population having greater between-subject variation than within subject variation. If within-subject variation exceeds between-subject variation then very low correlation values will be observed since the rank order of the data may change a great deal between repeats, despite the scores themselves only changing by a relatively small amount. It is therefore difficult to tell if a low correlation coefficient is indicative of a small amount of between-subject variation or poor replicability. Because of these limitations the correlation between repeats is not explored as an indicator of measurement precision for the CRM. Alternatively, the two measures of reliability described below avoid these problems.

Variability is a measure of the variation in a single subject's score across repeated measurements and can be described as the typical error of measurement. If a subject is tested repeatedly under the same test conditions then a test with low variability would display minimal change in scores between repeats. The variability between repeated scores is caused by random error (Summerfield et al, 1994) and is quantified by calculating the within-subject standard deviation, using one way analysis of variance, and presenting the 95% confidence limits of this value (Bland, 2000). This value can be used to show the 95% confidence limits within which a subject's true score will lie, based on a single measured score (Bland, 2000). This measure has been previously used for assessing the reliability of speech recognition tests (Summerfield et al, 1994). This value gives an indication of the typical error or measurement of a test. Providing there is no indication of any systematic error between repeats (such as a pattern of decreased performance caused by fatigue or increased performance caused by a learning effect) variability can be used to indicate the amount of random error with a measurement tool. Calculating variability across an entire sample assumes that the typical measurement error is the same across all the subjects (homoscedastic) and it is therefore acceptable to use one value to describe the entire sample. If data is heteroscedastic (contains sub-populations that have different magnitudes of typical measurement error) then it is sensible to assess variability for individual subgroups (such as the normal hearing and hearing impaired populations), rather than averaging across the sample.

Stability of test results is the final aspect of reliability assessment; this simply means the magnitude of the change in score over repeats. A test with good stability displays a small change in scores between repeated measures under identical test conditions. Stability is quantified by calculating the mean difference between repeated measures and reporting this value and the 95% confidence limit of the mean difference. Since the variability of a test describes the amount of random error it is possible to tell whether a change in the means between repeats is caused by a

systematic effect or by random error by comparing the stability and variability values. If the magnitude of change between repeats exceeds the typical error of measurement then this would be indicative of either a true change in an individual's test score or systematic error, such as learning effect or fatigue. If the typical error of measurement value is the same size or greater than the change in mean then this indicates that the change in mean across repeats can be accounted for by random error rather than systematic error (Summerfield et al, 1994).

The second aspect of measurement precision shown in Figure 5.13 is the validity of the measure. The term validity is often used to encompass a variety of different concepts (see Appendix A). When assessing measurement precision of a test, the type of validity in question is that which explores the level of agreement between a measured value and its 'true' value. The 'true' value is a hypothetical concept and immeasurable, so validity is quantified by comparing measurement values with values which are as close to the 'true' value as possible (Hopkins, 2000). The term '**concurrent validity**' is used to describe this. It refers to the level of agreement between scores on a particular test and scores on another test which is measuring the same ability (McLeod, 2007). Since it is not possible to measure a 'true value', such as an individual's 'true SRT', a second test which is known to measure the set of abilities of interest can be used as the closest method for testing this. Concurrent validity can then be assessed using two methods: 1) correlation and 2) similarity of test scores. The correlation between the scores on the test in question and a second test using simple correlation analysis provides an indication of the level of agreement between the two scores, which is expected to be high if they are in fact measuring the same ability (Hopkins, 2000). However, as discussed previously, correlation analysis does not give an indication of the similarity between measurement values. For this reason, the correlation analysis should be combined with an assessment of the similarity of scores between the test in question and the test measuring the same set of abilities. This can be carried out by calculating the mean difference between the scores. The concurrent validity of the CRM will be investigated by measuring the following:

1. Correlation between the CRM adaptive procedure and
    a. an alternative measure of SIN ability, the TDT
    b. an alternative measure of hearing acuity, PTA
2. Similarity of scores on the CRM and the TDT.

The alternative measure of SIN ability that has been chosen to measure the concurrent validity of the CRM is the TDT (Lutman et al, 2006).The TDT sentence format is *"The digits one, two, three"* for example, whereby three random digits (from zero, pronounced 'O', to ten, excluding disyllabic seven) are presented after the carrier phrase. The TDT has been selected for two reasons. Firstly it

is a widely used SIN test which is used as a screening tool to assess SIN ability, the same construct which the CRM has been designed to assess (Kinson, 2012). Secondly, the limited variation between the CRM and TDT test formats make it an appropriate choice for measuring concurrent validity since it is assumed that the two tests will be measuring the same auditory abilities. The similarities between the TDT and the CRM are: 1) a carrier phrase is used (*"The digits…"*); 2) it is closed set; 3) there is no syntactic or semantic information; 4) the same scoring method is used; 5) the same numbers of key words need to be recalled (when the subject is responding all three target words for the CRM); and 6) they are both measured in stationary speech-spectrum noise. Since it assumed that similar SIN abilities are being measured by the TDT and the CRM the scores on the TDT can be used to measure concurrent validity through correlation analysis and by investigating the similarity of the scores on the TDT with the CRM.

The TDT was originally developed in Dutch at the Vrije University in Amsterdam and was only accessible by telephone, as it was designed to be used as a telephone screening tool (Kinson, 2012). Due to the success of the screening tool in Amsterdam the test was converted into numerous other European languages, including English (Lutman et al, 2006). Data about the validation of the TDT in other languages has been reported (HearCom, 2006) but there is no published data about the validation of the English recording of the TDT. A Masters dissertation (Hall, 2006) reports the development of the English TDT but there is no published data about its validation. However, considering it is wide spread use as a screening measure by Action on Hearing Loss it is assumed that the measurement precision of the test has been assessed but not reported in the literature.

As part of measuring the concurrent validity of the CRM correlation analysis can be conducted between the CRM scores and pure-tone thresholds measured using PTA. It is expected that there will be a correlation between PTA and scores on the CRM. The CRM test is carried out suprathreshold, and therefore is not measuring audibility, the construct being measured during PTA. However, given that sensorineural hearing impairment not only affects audibility but also affects an individuals to process the sounds (Plomp, 1978, see Chapter 4 for further details) it is expected that as averaged pure-tone thresholds deteriorate so will CRM thresholds. Since PTA and the CRM are measured in different units (thresholds in dB HL and SRTs in dB SNR) only the correlation analysis aspect of concurrent validity can be investigated; calculating the similarity of test scores would be meaningless.

To summarise, the assessment of measurement precision can be split into two categories, reliability and concurrent validity. The aim of Study 4 is to investigate the various aspects of measurement precision for the CRM test, allowing for an informed decision to be made about

whether it is appropriate to further assess the CRM as a measure of AFFD. However, at this stage there are still some uncertainties about the best test conditions for the CRM adaptive procedure test. These can be broken down into two topics:

1. the adaptive procedure parameters
2. the scoring method for the CRM

The use of adaptive procedures for measure SRTs in noise is a very common methodology (Lutman et al, 2006; Leek, 2001; Summerfield et al, 1994; Ozimek et al, 2009). There are a number of variations of adaptive procedures, each with their own merits and faults (Levitt, 1971), which can be employed for measuring speech recognition. However, it is not the purpose of this study to carry out an in-depth analysis of adaptive procedure methodologies and therefore only one adaptive procedure will be explored. An explanation and justification for the chosen procedure is provided in Section 5.5.3.

Uncertainties relating to the best scoring method for the CRM adaptive procedure test will be further explored in Study 4. The scoring method which will provide the highest measurement precision is unknown. As discussed in Section 5.4.8, previous uses of the CRM have only required participants to respond to the colour and number aspect of the CRM; this is because the speech test is mainly used to measure informational masking or localisation and the call sign was being used as a target word for participants to lock onto (Eddins & Liu, 2012; Bolia et al, 2000; Brungart, 2001a; Brungart, 2001b; Rothpletz et al, 2011). The aim here to use the CRM as a measure of SIN ability; there is no need for a target phrase so it may be feasible to include the call signs as a word for identification or to not include it as a key word. In order to broaden the potential future uses of the CRM both scoring methods will be investigated at this stage. In terms of using the CRM as a measure of AFFD the scoring method that demonstrates the best measurement precision should be selected. Study 3 has shown the call sign target word group to be the least homogenous and to contain the words with the shallowest slopes; hence there is some concern that scoring responses to the call sign target word may reduce the overall measurement precision of the CRM adaptive procedure. For this reason, measurement precision will be investigated both with and without responses to the call sign being scored.

To summarise, there are a number of aspects of measurement precision which should all be assessed in conjunction to evaluate suitability of the CRM adaptive procedure as a measure of speech recognition in noise. The scoring method with the highest measurement precision is currently unknown and therefore the measurement precision of two different CRM test conditions will be investigated (responding to all three target words or only the colour and number target words). There is no definitive way of comparing measurement precision between

test conditions; it is a case of building up a picture for each condition and assessing whether any one test condition consistently outperforms the others in terms of measurement precision. The ultimate aim of Study 4 is to investigate the measurement precision values of the two CRM test conditions and to decide whether any one scoring method out performs the other. This information can then be used to make an informed decision about which test condition should be further investigated as a measure of AFFD.

5.5.2 Research objective 4 and Study 4 aims

Knowledge gap: The measurement precision of the two CRM adaptive procedure test conditions (two scoring methods) as measures of SIN ability is currently unknown. This information is required to make an informed decision about whether the CRM is a suitable SIN test to further investigate as a measure of military AFFD.

Research objective 4: To investigate and compare the measurement precision of the two CRM test adaptive procedure scoring methods and to investigate the concurrent validity of the test in comparison to the TDT, an alternative measure of SIN ability.

Aim 1: To measure the stability values of the two CRM scoring methods

Aim 2: To measure the variability of the two CRM scoring methods

Aim 3: To measure the concurrent validity of the two CRM scoring methods in comparison to an alternative measure of SIN ability

Aim 4: To decide which of the two CRM scoring methods should be further investigated for external validity by comparing: 1) the measurement precision values (stability, repeatability and concurrent validity) and 2) the difference between SRT scores for normal hearing and hearing impaired individuals, as defined by PTA.

5.5.3 CRM adaptive procedure characteristics

In order for the CRM to be used as a measure of AFFD within the Armed Forces it needs to be implemented in an adaptive procedure, a fast and efficient method for determining individual SRTs. In the most basic sense an adaptive procedure is a psychophysical method whereby the presented stimulus level is determined by the preceding stimuli level and response. Up-down methods are a subset of adaptive procedure methods and are known as sequential experiment procedures, whereby the course of the experiment is dependent on the experimental data (Levitt, 1971). As explained in Section 5.5.1 adaptive procedures allow for relevant observations about

the PF to be made with maximum efficiency but without sacrificing accuracy and the method is commonly used for measuring SIN thresholds (Leek, 2001). There are several different types of adaptive procedures which focus on providing information about different aspects of the PF; in the main these are the threshold, location and the slope of the function. The up-down adaptive procedure focuses on making relevant observations to measure the location of the PF (Kingdom & Prins, 2010, see Appendix B for an overview on PFs). There are a large number of variations of up/down methods. The main factors which change between methodologies are: 1) the up-down ratio; 2) the stimulus intensity step size rules; and 3) data analysis and reported score. The aim of the CRM adaptive procedure is to measure an individual's SRT. Each of these factors is considered in more detail below and a decision has been made about the chosen methodology for the CRM adaptive procedure test.

1. Up-down ratio

The response sequence rules refer to the number of correct/incorrect responses required for a change in stimulus intensity. The simple up-down procedure estimates the 50% correct level and follows the rule of decreasing the stimulus intensity (the SNR) after a positive response or increasing it following a negative response. Alternatively, the transformed two-down one-up procedure, introduced by Levitt (1971) requires two correct responses before the SNR is decreased and one incorrect response between the SNR is increased. There are two advantages to the transformed up-down method over the simple up-down procedure. Firstly, although this method requires more trials to reach near threshold it helps ensure the stimulus intensity is only made harder after a 'true' correct response, rather than after one correctly guessed response. This prevents the SNR being decreased to below threshold after a series of guesses. Secondly, there is evidence to suggest that the efficiency and reliability of an adaptive procedure increases rapidly when the number of responses required for a change in stimulus intensity is greater than one (Brand & Kollmeier, 2002). An implication of requiring more than two correct (e.g. a three-down one-up procedure) is that the more complicated the sequence rule the more trials required in an adaptive track, creating a more tedious procedure for participants (Leek, 2001). The weighted method used influences the target proportion correct point on the PF, focusing on the 70.7% correct point (Levitt, 1971). A two-down one-up response sequence rule will be used in the CRM adaptive procedure test.

2. Step size rules

The step size rules refer to the increments by which the stimulus are increased or decreased following a trial. In the event that little is known about the spread of location of the PF that is being measured it is desirable to start at a SNR at which it is guaranteed to elicit responses of near

100%. If the step sizes following this initial presentation are small this results in a large number of suprathreshold presentations which are not relevant for measuring the PF. Choosing a single step size for the entire duration of an adaptive procedure can be problematic since a small step size results in many wasted presentations prior to converging at the threshold point and a large step size makes it difficult to gather accurate threshold estimation (Levitt, 1971). The use of large step sizes at the start of the procedure allows for the targeted threshold to be reached early on in the run, avoiding numerous suprathreshold presentations. By decreasing the step sizes after the first few reversals this means the majority of presentations are focused at the threshold point. Using a very large step size at the start of the procedure allows for a very easy presentation level to be used at the start of the run, increasing the likelihood of a correct response for the first trial (Kingdom & Prins, 2010). The chosen step size rules are 8dB for the first reversal, 4dB for the second and third reversals and 2dB for the remaining eight reversals. This will quickly target the threshold level, whilst still having a starting presentation level well above the expect threshold, and then use smaller step sizes to gather a more accurate threshold estimate.

3. Termination criteria

The termination criteria refers to the rules applied which prevent the adaptive procedure from running indefinitely but rather finishes when some predetermined criteria have been reached. Several termination rules have been proposed but it is most common that the procedure is terminated after a specific number of reversals have occurred at a given step size (Quintana & Pérez, 2003). There is no set rule for the number of reversals which should be completed before a procedure is terminated. It is important that enough data is collected to provide a good estimate of the threshold value. Brand and Kollmeier (2002) evaluated the total number of sentences presentations (using German sentence tests) required before a change in stimulus intensity to obtain a reliable SRT (approximately 1dB standard deviation) is at least 30. For the CRM adaptive procedure the trials are terminated after eight reversals for the smallest step size. This number of reversals will ensure a minimum of 30 CRM sentences are presented. The procedure is also terminated if the participant has listened to 60 trials before they have completed eight reversals of the smallest step size. This prevents data from inconsistent listeners being gathered and stops the procedure from going on indefinitely.

4. Data analysis and reported score

The method used for data analysis for an adaptive procedure directly influences the reported score and there are several available methods. One method involves pooling all of the data from each presentation and using it to fit a PF, which can be used to read off the threshold value. However, this methodology requires that there has been no change in the parameter values

during testing, such as varying step size rules. An alternative and simpler method of estimation, developed by Wetherill (1963, cited in Levitt, 1971) involves taking the mean value of the reversal values. For procedures with varying step sizes the reversals for the smallest step sizes should be used since the presentations at this point will be focused around the threshold, therefore providing more accurate data for the threshold estimation (Kingdom & Prins, 2010). Using this method of estimation is the equivalent of taking the midpoint value of every second run as the estimated threshold. For the CRM adaptive procedure the mean SNR is taken for the final eight reversals, where the step size is smallest, providing the SRT estimate. Theoretically this equates to the 70.7% correct point, since this is the threshold targeted by a two-down one-up procedure (Levitt, 1971). Provided that the targeted percentage correct point is above the guess rate and below the lapse rate the specific value is of little relevance since all SRT scores are measured in the same way and are therefore directly comparable.

To summarise, there are a number of different types of adaptive procedure, each with their own merits and faults. The purpose of Study 4 is not to evaluate the best type of procedure to be used for running the CRM adaptive procedure test. Section 5.5.3 has outlined and justified the chosen adaptive procedure method, which is described in detail in Figure 5.14.

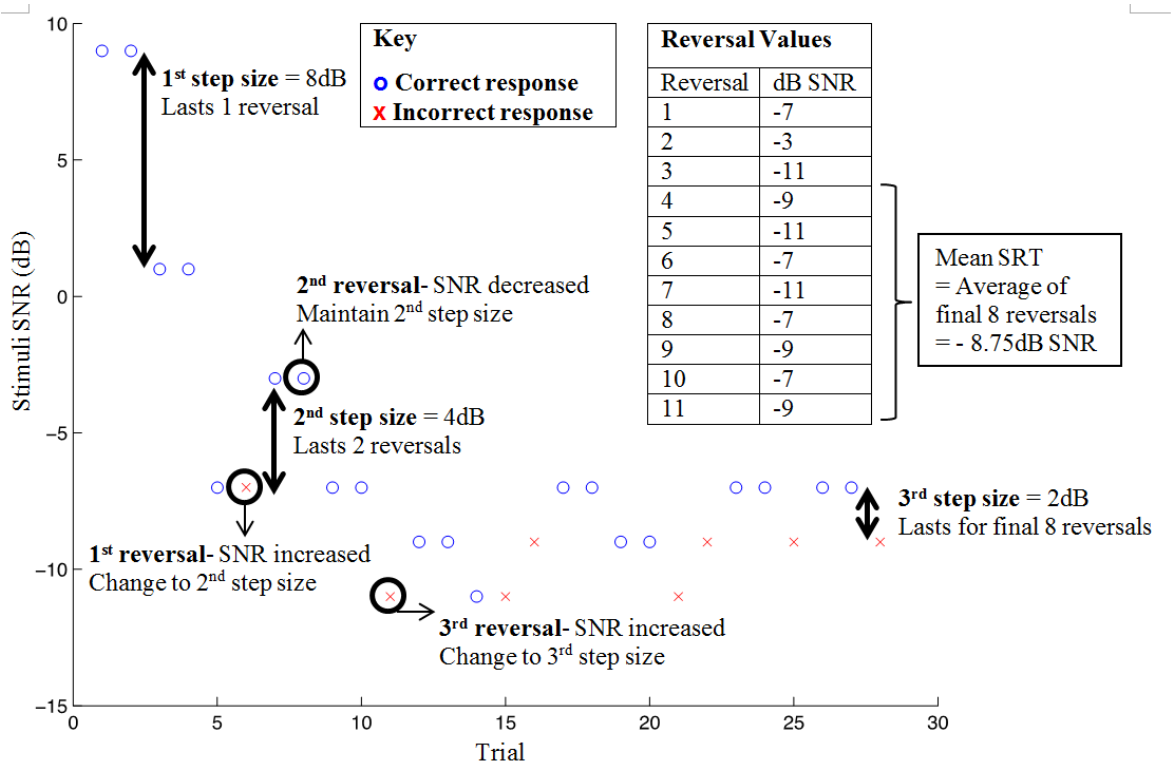


Figure 5.14 *Example adaptive procedure response plot with labelled key features. Example taken from a hearing impaired subject (better hearing ear average 31dB HL) listening to the CRM sentences in stationary speech-spectrum noise responding to the colour and number target words of the sentence.*

5.5.4 Methods

Normal hearing (NH) and hearing impaired (HI) listeners participated in the current study. The NH subjects were tested at the Institute of Sound and Vibration Research, Southampton, UK and the HI participants were tested at the Institute of Naval Medicine, Gosport, UK. All subjects were native speakers of the English language. Ethical approval was obtained for this study from the University of Southampton (ERGO ref: 13712) and MoD Research Ethical Committee (ref: 505).

The NH group were recruited from the University of Southampton. Individuals were invited to take part via email if they were an audiology undergraduate student or a postgraduate student within the Faculty of Environment and Engineering at The University of Southampton. The participant group consisted of 30 NH listeners. It was not appropriate to carry out a sample size calculation to address the main aim of this study, assessing the measurement precision of the CRM adaptive procedure. The chosen sample size was selected to obtain an estimate of the measurement precision values of the CRM. Table 5.8 contains further information about these participants. Normal hearing was defined as having hearing thresholds of ≤20dB HL at 0.25, 0.5, 1, 2, 3, 4, 6 and 8 kHz in both ears and being otherwise otologically normal (no tinnitus or current ear disease).

The HI group were recruited from the Defence Audiology Service at the Institute of Naval Medicine. The HI participant group consisted of 22 listeners (see Table 5.8). Initially the aim was to recruit 30 participants, to match the NH sample, but this was not possible within the time frame of the study (see Appendix O for details of recruitment challenges in the military population). Individuals were invited to take part via post if: 1) they were due to attend an appointment at the Defence Audiology Service within the time frame of the study and 2) their previous audiogram (completed within the last 12 months) indicated they had suitable hearing thresholds within the recruitment guidelines (a sensorineural hearing loss in their better hearing ear) with a mean thresholds of > 20 and < 70 dB HL averaged across 0.25, 0.5, 1, 2, 3, 4, 6 and 8 kHz). The spread of hearing abilities was dictated by the patient population attending appointments at the Defence Audiology Service; this resulted in more participants with milder hearing losses compared to those with more severe impairments (see Figure 5.15). Participants responded to the invitation by posting a consent form to the author and the testing session occurred immediately after they attended their audiology appointment at the Defence Audiology Service.

Table 5.8 *Gender and age sample characteristics of participants in Study 4*

	Normal hearing (n=30)	**Hearing impaired (n=22)**
Gender	Male n=10 Female n=20	Male n=20 Female n=2
Age (years)	Mean = 24 18-30 n=28 31-40 n=1 41-50 n=1 51-60 = n=0	Mean = 46 years 18–30 n=0 31-40 n=5 40-50 n=10 51-60 = n=7

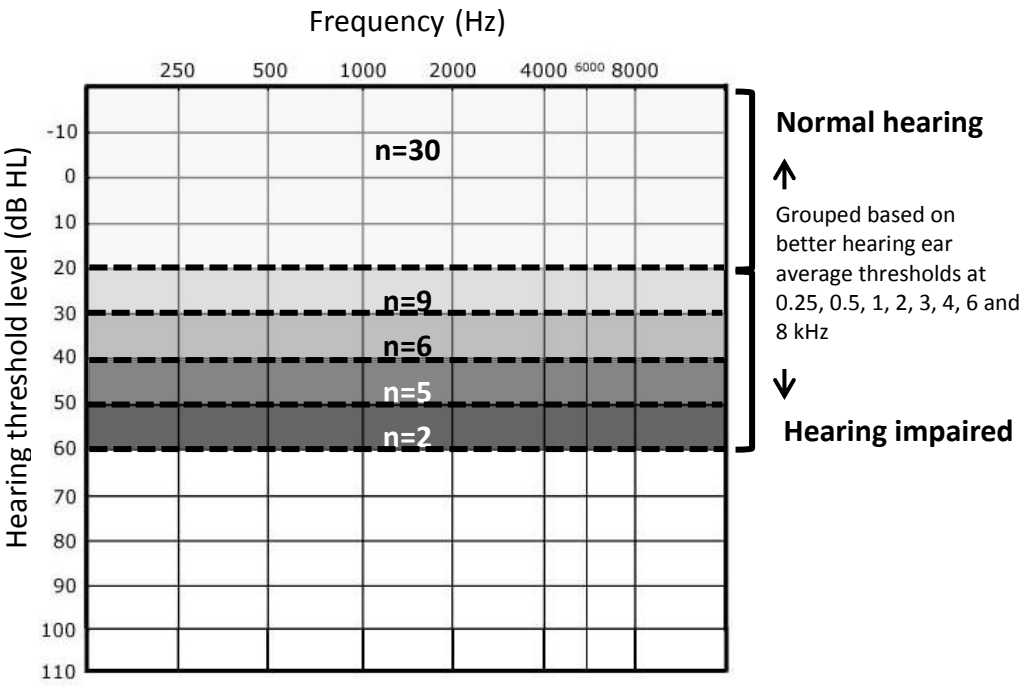


Figure 5.15 *Hearing acuity distribution of participants in Study 4*

For both the NH and HI groups participation was on a purely voluntary basis, no payment was provided. Participant information sheets were provided at least 24 hours prior to signing the consent form and data collection commencing.

Both the NH and HI listeners took part in SIN tests which employed an adaptive procedure to estimate their mean speech intelligibility threshold. In the adaptive procedure, the noise was kept at a constant level and the speech signal was varied. The test conditions were different for the NH and HI listeners, as detailed in Table 5.9.

The adaptive procedure rules were the same across all the conditions for both the NH and HI listeners. A two-down-one-up procedure was employed, the noise was kept at a constant level and the speech level was varied. The equation used to decide how much gain to apply to the sentence at each SNR is the same as that used in Study 3 (see Equation 6, p. 21). The procedure used three different step sizes, starting with a large step size of 8dB, for one reversal, then reducing this to 4dB for 2 reversals and then using 2dB steps for the remaining eight reversals. The adaptive procedure parameters are shown in Figure 5.14.

The CRM sentences were generated and scored using specifically designed MATLAB (R2013b) code. The experiment was run using a Mac laptop, running OS X Version 10.9.1. The stimuli were presented via an RME Babyface external sound card through Senheisser HDA 200 headphones. Calibration was performed in an artificial ear type 4153 using a flat plate coupler. A custom noise file was created for every sentence presented; this ensures the noise and sentence are the same length and that the noise file is not identical for each sentence. This custom noise file is a randomly selected segment from within a 28 second long stationary speech-spectrum noise file (see Appendix D to see how this was generated).

Both the NH and HI participants were given the same instructions. The following key points were communicated:

- You are going to listen to some sentences over headphones, with a background noise masker.
- You will have the following screen in front of you (the participant is shown the graphical user interface for the TDT and CRM, both with and without the call signs). Using the buttons, respond to what you have heard and when you are happy click to play the next sentence.
- When you begin it should be easy to hear the sentences. If you get a sentence correct the speech will get quieter until I expect that you will be unable to hear the sentence.
- If you do not know what you heard then simply guess. If you get a sentence wrong the speech will get louder, until you are able to respond correctly again.
- This process will continue for between 4-6 minutes until the test has finished.
- This test is measuring your threshold, the point at which you can just hear what is being said. For this reason you may feel as though you find the test quite difficult throughout. Do not worry, I expect this to be the case, it is not interesting for me to measure the points at which you get 100% or 0%.
- If you need a break at any time simply remove the headphones and don’t click to hear the next sentence.

Table 5.9 *Study 4 test conditions for normal hearing (NH) and hearing impaired (HI) participants*

Test	Background Noise	Call signs on or off[3]	Noise presentation Level (dB A)		Starting SNR (dB)		Maximum SNR (dB)[8]		Sessions and repeats[9]	
			NH[4]	HI[5]	NH[6]	HI[7]	NH	HI	NH	HI
CRM	Stationary speech-spectrum[1&2]	On	63	63-85	-1	9	10	10	Two sessions, two repeats per session (total four repeats)	One session, two repeats per session (total two repeat)
		Off								
TDT		NA			-5	5				

Table Notes

1 **CRM stationary speech-spectrum noise:** The speech-spectrum noise is the same as that described in Appendix D

2 **TDT stationary speech-spectrum noise:** The same method as described in Appendix D was used to create a white noise wav file with the same frequency shaping as the TDT sentences.

3 **Call signs on or off:** When the call signs were 'on' participants had to respond to all three target words of the CRM sentence (call sign, colour and number). When the call signs were 'off' participants only responded to the final two CRM target words (colour and number).

4 **Normal hearing noise presentation level:** The noise was presented at a constant level 63dB A which was the highest level for safe noise exposure for the test duration. Smits et al (2004) reported that so long as the presentation level of a speech test is audible and comfortable the specific presentation level does not affect threshold. The chosen presentation level is the same as that used in Study 3, which was based on that used by Ozimek et al (2009), in a similar study developing the Polish TDT.

5 **Hearing impaired noise presentation level:** This level varied between participants. The lower limit was 63dB A, the same as the presentation level used for normal hearing adults. The upper level was 85dB A which was selected because it was the limit for noise exposure for normal hearing listeners and therefore prevented any ethical issues arising from exceeding the daily noise dosage (Human Experimentation Safety and Ethics Committee, 1996). Participants were played CRM and TDT sentences (separately) and where asked to adjust the volume using the dial on the SLM to a level that was comfortable for them to understand the speech. They were all given the same instructions: 1) you should not be straining to understand the speech; 2) it should not be uncomfortably loud; 3) if you wear hearing aids it should be a similar a volume to listening to speech in quiet through your hearing aids. This ensured that the starting presentation level would present the speech at a level at which they should score 100% correct.

6 **Justification for normal hearing starting SNR:** This value was selected from the Study 3 results; at -1dB SNR participants scored 100% correct when responding to the whole CRM sentence.

7 **Justification for hearing impaired starting SNR:** Leensen et al (2011b) showed that hearing impaired listeners needed around a 5dB more advantageous SNR compared to normal hearing listeners when listening to speech in speech-spectrum background noise to achieve the same SRT. The hearing impaired sample in their research had milder hearing losses than some of those involved in Study 4 so it was decided to provide a 10dB more advantageous SNR (starting at 9dB SNR) for the hearing impaired listeners to ensure that the first presentation was at a level which should be easy to hear and elicit a 100% correct score.

8 **Justification for normal and hearing impaired maximum SNR**: To avoid the presentation level reaching a very high SNR as a result of a series of suprathreshold incorrect responses a cut-off of 10dB SNR has been implemented.

9 **Sessions and repeats**: The normal hearing sample attended two sessions and at each session their SRT was measured twice for each test condition, resulting in four repeated measurements. The hearing impaired sample could only attend one session, resulting in two SRT for each test condition. The hearing impaired participants took part in the study immediately following their audiology appointment at the Defence Audiology Service; because participants travelled from all over the UK for these appointments it was not possible to ask them to return for a second session as this would incur significant travel costs.

Data was analysed using IBM SPSS Statistical Analysis Software (version 19).Both descriptive and inferential statistical analysis will be carried out. Prior to conducting any inferential statistical testing the data will be checked for normal distribution using the Shapiro-Wilk test, which is appropriate for small sample sizes (Field, 2005a). The results of this test will be used to ensure appropriate parametric or non-parametric statistical tests are employed. The following bullet points indicate the fundamental methods that will be used to explore each aim.

- **Aim 1: stability**. The mean change in scores between repeats and sessions (for the NH sample only) will be calculated. Two repeated-measures ANOVAs will be conducted (for the NH and HI samples separately) to explore the main effects of condition (the two CRM scoring methods and the TDT), repeat and sessions (for the NH sample only).
- **Aim 2: variability**. The 95% confidence interval of the true SRT score for any one measurement value will be calculated using the within-subject-variation across repeats, providing an estimate of typical error of measurement.
- **Aim 3**: **concurrent validity**. Correlation coefficients between the TDT and the two CRM conditions and PTA and CRM conditions will be calculated. The similarity of scores (looking at the mean difference in SRTs) between the TDT and the CRM test conditions will be calculated.
- **Aim 4**: **comparing test conditions.** Comparisons of the results from the aforementioned tests will be made to assess whether any one scoring method particular test outperforms the other in terms of measurement precision. If no difference is found between the scoring methods then a correlation coefficient comparing results on the two scoring methods will be carried out; if the two test conditions are correlated with each other statistically it would not matter which condition was chosen to be used as a measure of AFFD; the SRT for any one CRM test condition could be used to predict the SRT for the other CRM test condition.

5.5.5 Results: overview

The abbreviations detailed in the final column of Table 5.10 will be used throughout the remainder of Chapter 5 to describe the test conditions experienced by the normal hearing (NH) and hearing impaired (HI) participant grounds. Table 5.9 refers to the methodologies for these conditions.

Table 5.10 *Study 4 CRM test condition abbreviations*

Stimuli	Call sign on/off	Abbreviation
Coordinate Response Measure	Call sign on	CRM-CSon
	Call sign off	CRM-CSoff
Triple Digit Test	NA	TDT

The results of Study 4 are reported across sections 5.5.5 to 5.5.9. This section (5.5.5) provides an overview of the mean SRT, averaged across repeats for different levels of hearing acuity and also explores whether the data is normally distributed, which will influence the appropriate statistical methods to apply. Sections 5.5.6-5.5.8 explores the aspects of measurement precision outlined in Figure 5.13. Finally, in Section 5.5.9 comparisons are made between the five test conditions.

The box plots in Figure 5.16 show the distribution of the SRTs for three different levels of hearing acuity. The hearing acuity groups have been chosen for illustrative purposes only, to demonstrate the change in SRT as hearing impairment worsens. Participants results were placed in one of three hearing acuity groups based on their pure-tone thresholds averaged across eight frequencies (0.25, 0.5, 1, 2, 3, 4, 6 and 8 kHz). The descriptors normal, mild and moderate to severe have been added for ease of reporting results.

- Normal hearing (n=30) averaged pure-tone thresholds ≤ 20 dB HL
- Mild hearing loss (n=12) averaged pure-tone thresholds >20 dB HL and ≤35 dB HL
- Moderate to severe hearing loss (n=10) averaged pure-tone thresholds >35 dB HL and ≤ 55 dB HL

For both the CRM conditions as hearing acuity worsens so do SRTs and the same pattern is observed for the TDT. Across all the groups participants achieve better SRTs when responding to the only the colour and number part of the sentence in comparison to responding to all three key words.

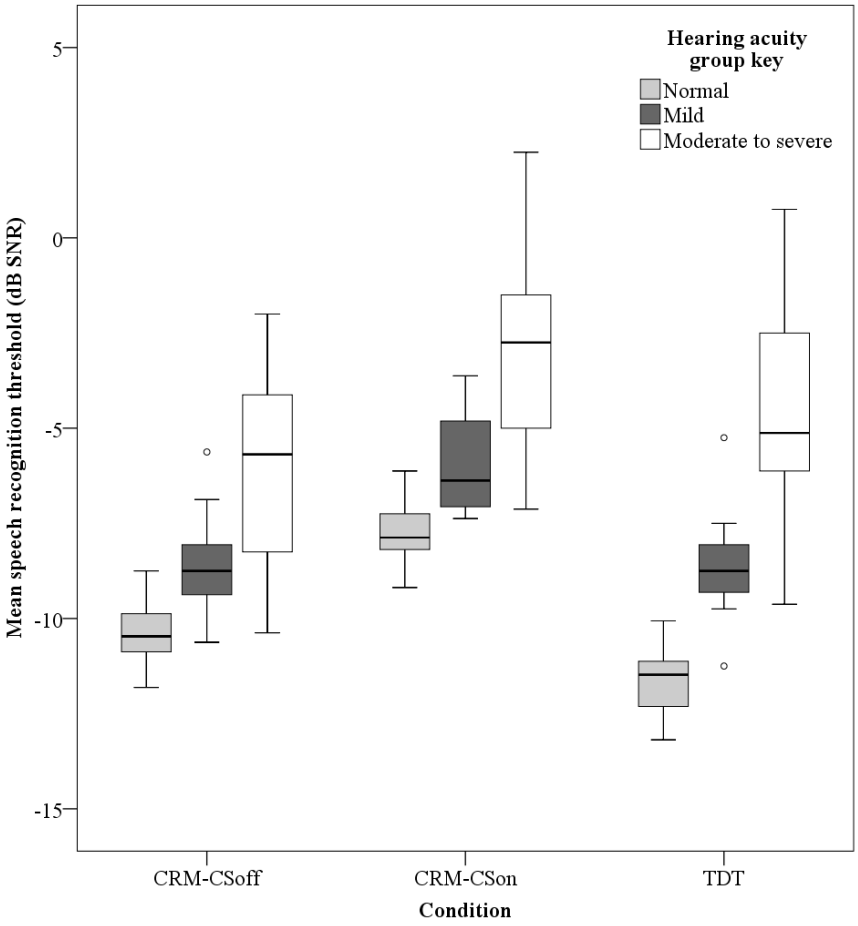


Figure 5.16 *Box plots showing the distribution of the SRTs for different levels of hearing acuity for each CRM conditions and the TDT (outliers exceed 1.5 times the interquartile range)*

Prior to conducting any further statistical analysis of the data normality testing has been carried out. The assessment for normally distributed data is a prerequisite for parametric statistical testing (Field, 2005a). The Shapiro-Wilk test is appropriate for small sample sizes (<50 samples) and it tests the null hypothesis that the data is normally distributed (Field, 2005a). The results for all the possible data sets which are used in the remaining results sections are tested for normality and the results are reported in Table 5.11.

The majority of the data sets for separately analysed normal hearing and hearing impaired samples are normally distributed. For the data sets which are not normally distributed there is no known reason why the distribution of this data should differ from the other data sets. There is no consistent pattern which indicates that a certain condition or repeat is generating non-normally distributed data. It is possible that that relatively small sample size has caused this result. A visual inspection of the histograms for these data sets did not show a distribution which differed a great deal in comparison to the other data sets. For these reasons it has been decided to treat all of the normal hearing and hearing impaired data as normally distributed and parametric testing has been carried out.

The combined normal hearing and hearing impaired data sets were all shown to be non-normally distributed. This result is not surprising since there are more normal hearing than hearing impaired participants, resulting in positively skewed data. Any analysis which is carried out which uses the combined data will use non-parametric testing methods.

Table 5.11 *Shapiro-Wilk test of normality across all the data*

Condition	Repeat	Shapiro-Wilk normality test ✓ = p>.05 or ✗ = p<.05		
		Normal Hearing	Hearing Impaired	Combined normal hearing and hearing impaired data
CRM-CSon	1	✓	✓	✗ (p<.001)
	2	✓	✓	✗ (p<.001)
	3	✗ (p=.007)	NA	NA
	4	✗ (p=.040)	NA	NA
	Mean across repeats	✓	✓	✗ (p<.001)
CRM-CSoff	1	✓	✓	✗ (p<.001)
	2	✓	✓	✗ (p<.001)
	3	✗ (p=.038)	NA	NA
	4	✗ (p=.007)	NA	NA
	Mean across repeats	✓	✓	✗ (p<.001)
TDT	1	✓	✓	✗ (p<.001)
	2	✓	✗ (p=.003)	✗ (p<.001)
	3	✓	NA	NA
	4	✓	NA	NA
	Mean across repeats	✓	✓	✗ (p<.001)

5.5.6 Results: stability

The first area of measurement precision to be investigated is the stability of the SRT scores across repeats. Stability has been assessed separately for the normal hearing and hearing impaired samples because the number of repeats was unequal for these two groups.

The stability analysis has been carried out based on the assumption that stability does not vary within the population, i.e. the magnitude of change in score between repeats is independent of SRT. To assess whether this is a fair assumption a set of Bland and Altman plots have been created for the normal hearing and impaired data sets to look at the within-subject variation between repeats and between sessions. These plots and a description of how they were created can be found in Appendix H. If stability was not affected by hearing ability then a visual inspection of the Bland and Altman plots would display no pattern in the differences in scores between

repeats/session across the samples. The plots in Appendix H display a fairly even distribution in the difference in SRT scores between repeats/session. This provides evidence that, for this sample, there is no reason to believe that the stability value may vary across either the normal hearing or hearing impaired samples. It is not known if this assumption can be made for the general population, this is explored further in the Discussion (Section 5.5.10).

Firstly, the stability across the four repeats, measured across two sessions (two repeats per session), for the normal hearing data will be explored. Figure 5.17 shows the mean SRT for each repeat, across the four conditions, with error bars displaying the 95% confidence interval of the mean. The mean changes in the SRT between repeats and sessions are outlined in Table 5.12. For all of the test conditions the mean within session change in SRT scores between repeats is less than 1dB. All conditions show an improvement in SRT scores between the first and second repeat within a session, apart from CRM-CSoff, which showed a small decrease (0.1 dB ± 0.4) in SRT score between the repeats in session two. The changes in SRT for the between session repeats was also small (1dB or less) for three of the conditions.

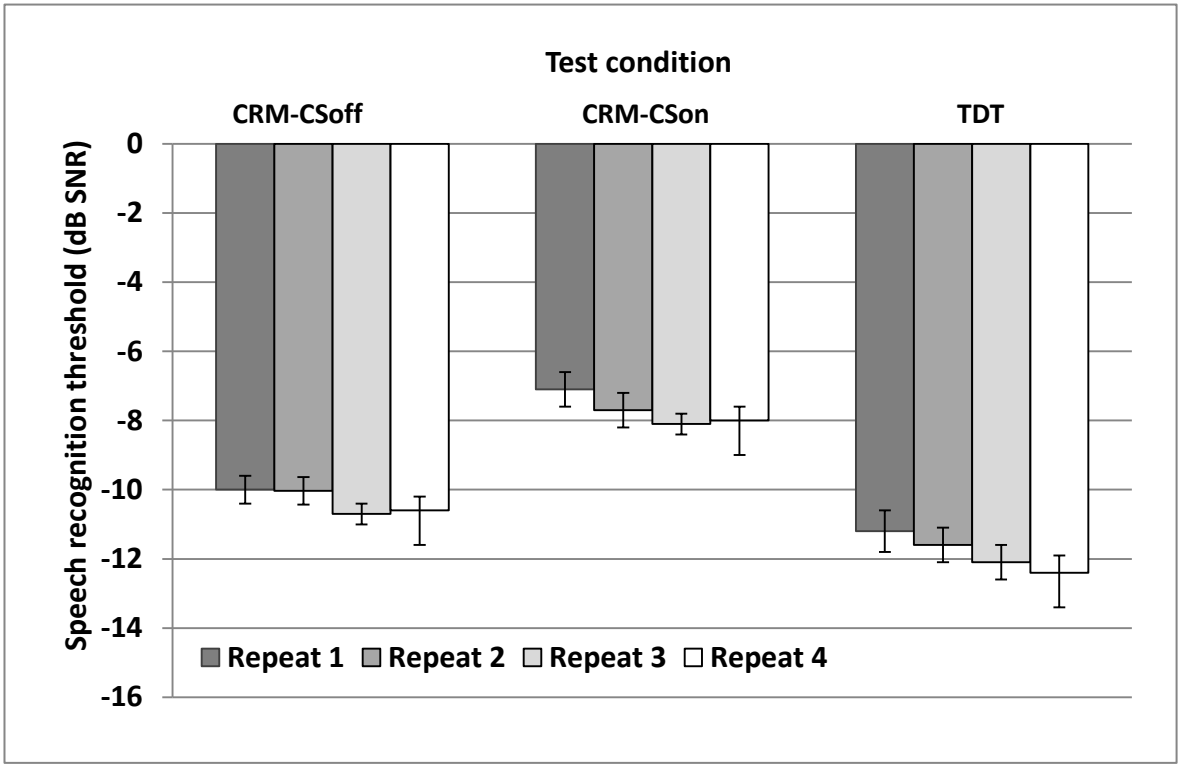


Figure 5.17 *Normal hearing sample mean SRTs for each repeat across the three test conditions. Error bars display 95% confidence interval of the mean.*

Table 5.12 *Normal hearing stability values for CRM adaptive procedure test conditions and TDT, showing the mean change in SRT between repeats and the 95% confidence interval of the mean.*

Condition	**Between repeats** ***(Session 1: repeat 1 minus repeat 2)*** **(dB)**	**Between repeats** ***(Session 2: repeat 3 minus repeat 4)*** **(dB)**	**Between session** ***(Repeat 1 minus repeat 3)*** **(dB)**
CRM-CSoff	0.3 (± 0.5)	-0.1 (± 0.4)	0.7 (± 0.5)
CRM-CSon	0.6 (± 0.5)	0.0 (± 0.5)	1.0 (± 0.6)
TDT	0.4 (± 0.6)	0.2 (± 0.6)	0.9 (± 0.6)

A three-way repeated measure ANOVA with condition, repeats and test conditions as the independent variables was conducted; the results are reported in Table 5.13. The key results from this ANOVA are reported below.

- There is a main effect of condition, indicating that averaged across all repeats SRT scores are significantly different between each condition. This difference is stable between sessions and repeats.
- Post hoc tests (paired sample T-tests) revealed a significant difference in the mean SRT scores (averaged across repeats) between each condition.
- There was a main effect of session, indicating that averaged across all conditions SRT scores are significantly better in session two than in session one. Post hoc tests (paired sample T-tests) revealed that between session means were significantly different for all test conditions ($p < .05$), however the change in SRT scores across means was small (< 0.7 dB).
- There was no main effect of repeats, indicating that, averaged across all conditions, SRT scores are not significantly different for the first within session SRT measurement compared to the second.
- The session*repeat interaction was significant, suggesting that the change in SRT scores between repeats was different within each session. Although there was a significant interaction the magnitude of any changes in the SRT scores either between repeats or between sessions the change in SRT score was small (< 1dB, see Table 5.12). It is to be expected that there is some learning effect across repeats and sessions, but the magnitude of this improvement is small.

Table 5.13 *Normal hearing three-way (condition, session and repeat) repeated measures ANOVA results*

Independent variables	Main effect	Interaction		
		Condition[1]	**Session[2]**	**Repeat[3]**
Condition[1]	*** F(2, 58) = 316.1 p < .001		F(2, 58) = 0.232 p > .05	F(2,58) = 0.501 p > 0.05)
Session[2]	*** F(1,29) = 14.91 p < .05			*** F(1,29) = 7.463 p < .05
Repeat[3]	F(1,29) = 3.754 p >.05			

Table notes

1 Three levels: CRM-CSoff, CRM-CSon and TDT

2 Two levels: session one, session two

3 Two levels: repeat one and repeat two, referring to the first and second repeat within each session

Secondly, the stability across the two repeats, measured within one session, for the hearing impaired data will be explored. Figure 5.18 shows the mean SRT for each repeat, across the five conditions, with error bars displaying the 95% confidence interval of the mean. It is understood that there is a great deal of variation in the SRTs amongst the hearing impaired samples which is not displayed in Figure 5.18 but the figure does provide an overall indication of the average change between the repeated SRT measurements.

The mean changes in the SRT between repeats are outlined in Table 5.14. For all the test conditions, the mean change between repeats is small (<0.5 dB). For all the CRM test conditions SRT scores improved from repeat one to repeat two however for the TDT on average the SRT scores worsened between repeats (-0.5 dB ± 1.3).

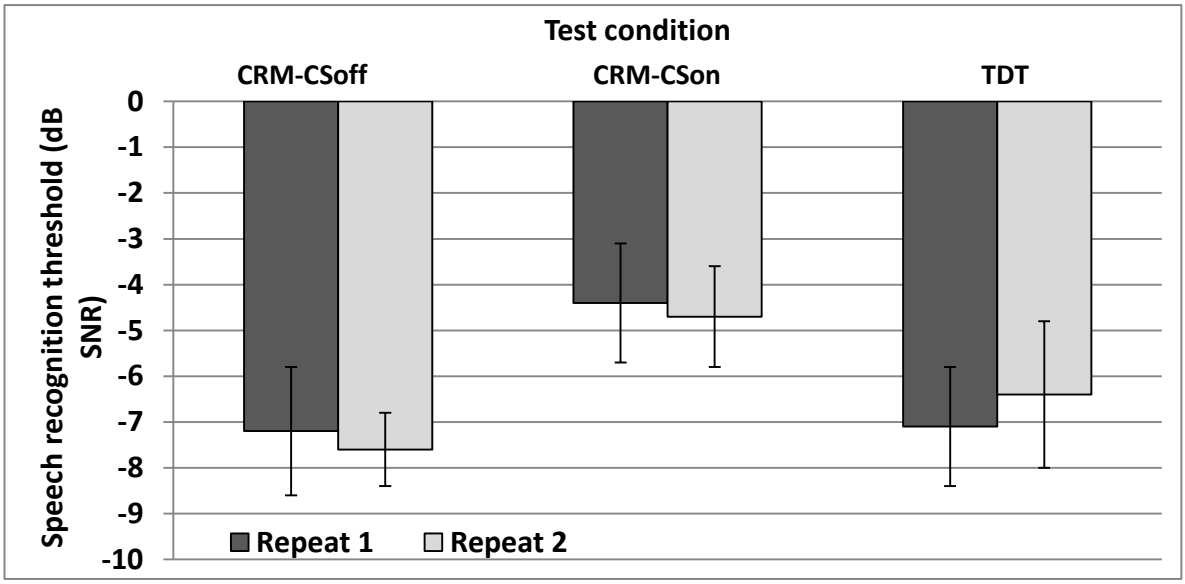


Figure 5.18 *Hearing impaired sample mean SRTs for each repeat across four test conditions. Error bars display 95% confidence interval of the mean.*

Table 5.14 *Hearing impaired stability values for CRM adaptive procedure test conditions and TDT, showing the mean change in SRT between repeats and the 95% confidence interval of the mean.*

Condition	Repeat 1 minus repeat 2 (dB)
CRM-CSoff	0.4 (± 0.8)
CRM-CSon	0.3 (± 0.6)
TDT	-0.5 (± 1.3)

A two-way repeated measure ANOVA with condition and repeat as the independent variables was conducted and the results are reported in Table 5.15.

- There is a main effect of condition, indicating that averaged across all repeats SRT scores are significantly different between each condition. This difference is stable between repeats.
- Post hoc tests (paired sample T-tests) revealed significant differences between two out of three pairs of conditions. The CRM-CSoff (M = −7.4 dB SNR, SD = 2.4) and TDT (M = −6.9 dB SNR, SD = 3.0) showed similar mean SRT scores; t (21) = −1.0, p = .3.
- There is no main effect of repeat indicating that the change in SRT scores across repeats, when averaged across all conditions, was not significant.

Table 5.15 *Hearing impaired two-way (condition and session) repeated measures ANOVA results*

Independent variables	**Main effect**	**Interaction**	
		Condition[1]	**Repeat[2]**
Condition[1]	*** F(1.3, 26.7) = 28.6 p < .001		F(2, 2.5) = 1.1 p > .05
Repeat[2]	F(1,21) = 0.29 p > 0.05)		

Table notes
1 Three levels: CRM-CSoff, CRM-CSon and TDT
2 Two levels: repeat one, repeat two

To summarise, it can be concluded that both the NH and HI results are stable across repeats and session. Across all test conditions a maximum change of 0.6 dB (± 0.5) was observed for within session measurements. Across the entire sample no significant difference was found between SRTs measured in the same session. For the NH sample SRT scores were found to be significantly better in session two than session one but the change was very small (≤ 0.7, ± 0.5).

5.5.7 Results: variability

The second aspect of measurement precision to be explored is the variability of results. This is displayed as the 95% confidence interval of the true SRT score for any one measurement value and is calculated by using the WSV across repeats to provide an estimate of the typical error of measurement for each test condition. Variability has been calculated separately for the normal hearing and hearing impaired samples because the data have unequal numbers of repeats.

When calculating variability an assumption is made that the data is homoscedastic (see Section 5.5.1). Sample homogeneity is important to ensure that any calculations of variability are applicable to the entire sample and do not overestimate or underestimate variability for sub-populations within a sample. Sample homogeneity is typically measured using Levene's test. However, with small sample sizes it is inevitable that there will be gaps in the distribution of the data, resulting in Levene's test indicating the data is heteroscedastic. As an alternative approach, the Bland and Altman plots in Appendix H can be used to investigate homogeneity of variance across the samples, between repeats and sessions. Homoscedastic data will display no pattern in the differences in score between repeats/session across the sample. If the data was heteroscedastic there would be a trend in the differences between repeats changing non-randomly across the sample.

As discussed in Section 5.5.6 there is no pattern in the differences in score between repeats/sessions across the sample for the normal hearing or hearing impaired data. For the sake of the analysis of variability the hearing impaired data set is assumed to be homoscedastic and since comparisons are being made between the test conditions which were measured using the same sample it is concluded that this is an acceptable analysis method. As is the case for the stability analysis, it is not known if this assumption can be made for the general population (explored further in the Discussion, Section 5.5.10).

Variability has been calculated by looking at the variation across repeated measurements for individual subjects. This involves taking the within-subject variance across four repeats for the NH sample and across two repeats for the HI sample. The within-subject variance is calculated by performing a one way ANOVA with the subjects as 'groups' and taking the square root of the residual mean square value (an estimate of how much variation in the data is caused by typical error of measurement) to give one standard deviation of SRT scores across repeats (Bland, 2000, p.269). To report this as a 95% confidence interval the standard deviation is multiplied by a standard normal distribution, 1.96 (Bland, 2000, p.270). This 95% confidence interval shows the confidence limits of the true SRT score for any one SRT measurement value. This calculation is showed in Equation 7. These values are reported in Table 5.16 for both the NH and HI samples.

Equation 7 *Typical error of measurement calculation*

$$Typical\ measurement\ error\ (95\%\ confidence\ interval) = 1.96 \times \sqrt{Residual\ Mean\ Square}$$

Table 5.16 *Variability (shown as the 95% confidence interval of the true SRT score for any one measurement value) for the CRM adaptive procedure conditions and the TDT*

Condition	Variability, 95% confidence interval (dB)	
	Normal hearing (across 4 repeats)	Hearing Impaired (across 2 repeats)
CRM-CSoff	1.9	2.5
CRM-CSon	2.1	1.7
TDT	2.6	3.9

In summary, for all of the test conditions the typical error of measurement (the 95% confidence interval reported in Table 5.16) is larger than the mean changes across repeats, indicating the changes across repeats can to some extent be accounted for by measurement error as opposed to being cause by a systematic change, such as learning effect.

5.5.8 Results: concurrent validity

The final aspect of measurement precision to be investigated is the concurrent validity of the CRM. There are two aspects of concurrent validity that will be explored in Section 5.5.8. Firstly, the correlation between scores on the CRM test conditions and both the TDT and the participants pure-tone thresholds will be explored. Secondly the similarities of scores on the CRM test conditions and the TDT will be investigated. It should however be noted that another aspect of concurrent validity is the comparisons of the measurement precision values between CRM test conditions and the TDT, reported in the previous two sections (5.5.6 and 5.5.7); this will be further explored in the Discussion Section (5.5.10).

Firstly the correlation aspect of concurrent validity is explored. Figures 5.19 and 5.20 are scatterplots which show the relationship between scores on the CRM test conditions and the TDT, averaged across repeats. Across both the CRM test conditions there is a positive correlation with the TDT scores. An initial visual inspection of the scatterplots does not indicate that any one CRM test condition is more or less correlated with the TDT.

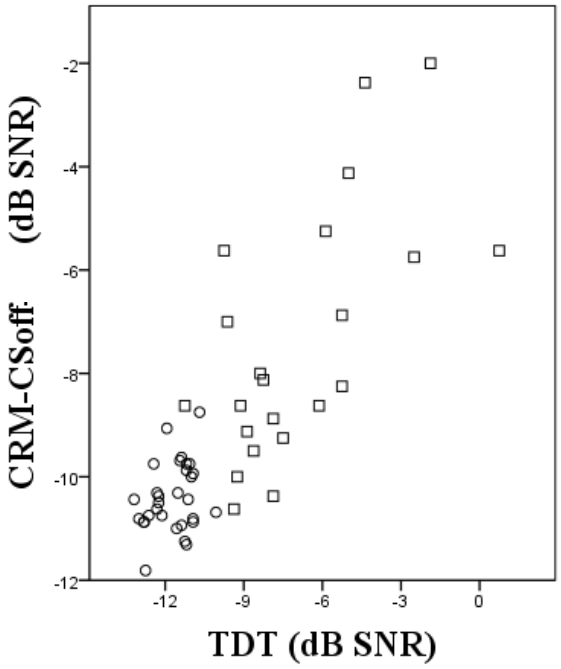


Figure 5.19 *Correlations between CRM-CSoff and TDT for normal hearing and hearing impaired data. Normal hearing =◯ and hearing impaired = □*

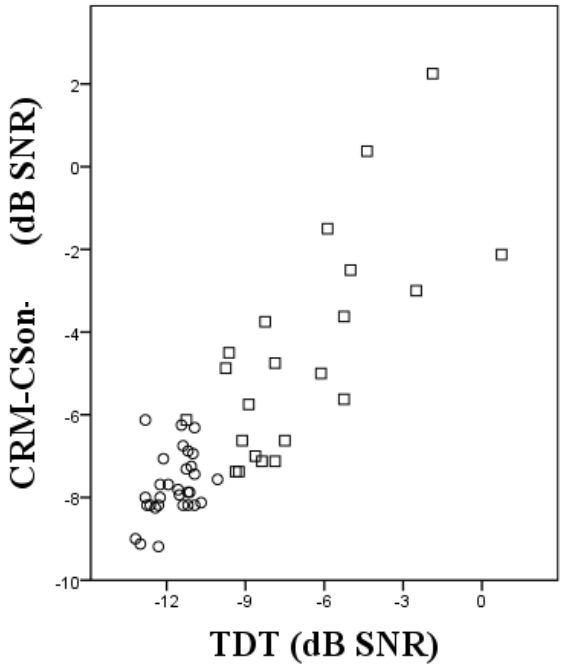


Figure 5.20 *Correlations between CRM-CSon and TDT for normal hearing and hearing impaired data. Normal hearing =◯ and hearing impaired = □*

Table 5.17 shows the correlation coefficients between the CRM test conditions and the TDT. Pearson's parametric correlation test (r) has been calculated for the normally distributed hearing impaired data and the non-parametric correlation test Spearman's rho (r_s) has been calculated for the combined normal hearing and hearing impaired data (see Section 5.5.5, Table 5.11 for Shapiro-Wilk normality test results).

All the correlations reported in Table 5.17 are significant ($p < .001$). The upper and lower correlation 95% confidence limits have been calculated using the equation outlined in Altman et al (2000, p.89), detailed in Equations 1-5 (see Section 4.4).

Table 5.17 *Concurrent validity between the CRM test conditions and the TDT. Pearson's parametric correlation test (r) has been calculated for the hearing impaired data. Spearman's rho (r_s) has been calculated for the normal hearing and combined normal hearing and hearing impaired data.*

Condition	Sample size, NH/HI	Correlation between CRM test conditions and TDT (SRT averaged across repeats)	Correlation 95% confidence limits, lower – upper (r)	r^2 (%)
CRM-CSoff	n = 22 (HI)	r = 0.65	0.31 – 0.84	42
	n = 52 (NH & HI)	r_s = 0.68	0.50 – 0.80	46
CRM-CSon	n = 22 (HI)	r = 0.72	0.43 – 0.88	52
	n = 52 (NH & HI)	r_s = 0.78	0.64 – 0.87	61

For both the CRM test conditions the effect sizes can be described as 'large' $(r \geq 0.5)$ according to Cohen's descriptions (Cohen 1988 and 1992, cited in Field, 2005a). The r^2 value gives an indication of what percentage of the variance in the CRM test condition scores is shared with the variance in the TDT scores. The CRM-CSon test condition has the best correlation with the TDT (hearing impaired data $r = 0.72$ and combined normal hearing and hearing impaired data $r_s = 0.78$). To assess whether there is a statistically significant difference in the concurrent validity between the two CRM scoring methods and TDT, the equality between the two correlations has been calculated using online software by Lee and Preacher (2013), which is based on the equations outlined in Steiger (1980) for comparing correlations.

Three values are required to values are require to calculate correlation comparisons; the two correlation coefficients to be compared (CRM-CSoff/TDT and CRM-CSon/TDT) and the correlation between the unshared variable (e.g. CRM-CSoff/CRM-CSon). A Z-score is given for each correlation comparison; by convention, Z-score values greater than 1.96 are considered significant if a two-tailed test is performed (Lee and Preacher, 2013). Correlation comparisons have been calculated separately for the combined NH and HI data and for the HI data. When comparing the correlation coefficients of the two CRM test conditions with the TDT, a Z-score of

<1.96 was obtained for both the data sets. It is therefore not suspected that any one CRM test condition is significantly more correlated with the TDT than any other, suggesting equal concurrent validity between the test conditions. This finding is consistent with the correlation 95% confidence intervals reported in Table 5.17 since all of these overlap.

The correlation between SRTs and the participant's pure-tone thresholds measured through PTA is another aspect of concurrent validity. It is expected that as hearing acuity, as measured by PTA, worsens so will SRTs. Table 5.18 shows the correlation coefficients between the speech test scores and the PTA results. The PTA score used is the better hearing ear averaged thresholds across eight audiometric frequencies, 0.25, 0.5, 1, 2, 3, 4, 6 and 8 kHz. This aspect of concurrent validity has only been investigated for the hearing impaired data since the normal hearing sample did not undergo a full audiogram but only a screening test, hence better hearing ear averages are not available. The same method used to compare CRM test condition correlations with the TDT has been used to compare the correlations between the SIN tests and PTA (Lee & Preacher, 2013). Given the similarities between the correlations in Table 5.18 it is not surprising that no one SIN test was found to be significantly more or less correlated with better hearing ear pure-tone threshold (Z < 1.96, Lee & Preacher, 2013).

Table 5.18 *Concurrent validity between the CRM test conditions (SRT averaged across repeats) and PTA (average pure-tone thresholds for better hearing ear) for the hearing impaired data, (Pearson's r).*

Condition	Hearing impaired		
	Correlation between CRM test conditions and PTA (r)	Correlation 95% confidence limits, lower – upper (r)	r^2 (%)
CRM-CSoff	0.74	0.46 - 0.89	55
CRM-CSon	0.75	0.48 - 0.89	56
TDT	0.80	0.57 - 0.91	64

Interestingly it appears that the CRM conditions are better correlated with PTA than with the TDT. This finding was unexpected since the CRM and TDT are both assessing similar auditory abilities required to understand SIN whereas PTA is only measuring the audibility aspect of hearing. To investigate this further the correlation comparisons have been carried out for each CRM test condition, looking at whether there is a significant difference between the CRM test condition correlation with the TDT and with PTA. These comparisons have only been carried out with the HI data as a better hearing ear average was not available for the NH sample and are shown in Table 5.19. It can be concluded that although the correlations coefficients between the CRM and PTA are higher than those with the TDT there does not appear to be a significant difference between the correlations and this finding may be due to small sample sizes or a chance finding from this particular data set.

Table 5.19 *Comparison of correlations of CRM test conditions with the TDT and PTA*

Condition	Correlation to PTA (r)	Correlation to TDT (r)	Is there is a significant difference between the CRM test condition correlation with the TDT and with PTA?
CRM-CSoff	0.74	0.65	No (Z= 0.92)
CRM-CSon	0.75	0.72	No (Z= 0.35)

The final element of concurrent validity to be explored is the similarity of the test scores. Following the finding that there is a correlation between scores on the CRM and the TDT it follows that if the TDT and the CRM are measuring the same or a similar set of auditory abilities then the scores on the two tests should be comparable. Prior to looking at the similarity in test scores between the two SIN measures it should first be checked that the TDT scores measured in this study match those found in the literature, providing evidence that the data collected in Study 4 is comparable to previous literature using the TDT.

Four main studies have been identified which contain values for the TDT for normal hearing listeners. It is simpler to compare the test with a normal hearing sample; comparing averaged TDT scores for hearing impaired samples would be problematic if the two groups contained very different hearing loss configurations. A literature search revealed only one study, part of a Masters Dissertation, which measured normal hearing performance on the TDT with the British stimuli (Hall, 2006). Table 5.20 provides information about the SRT thresholds on the TDT from a number of studies, both in English and other languages. The values reported in Table 5.20 for languages other than English have been taken from Ozimek et al (2009 p. 315, Figure 6) which looks at the SRT for the TDT across different languages, including English. The recorded TDT mean SRT from Study 4 is −11.8 (±) only differs from that recorded by Hall (2006) by 1 dB. The similarity between the SRTs measured in Study 4 and those found in the literature provides evidence that the TDT scores in Study 4 are reasonable.

Table 5.20 *Comparing the SRT scores for normal hearing listeners on the TDT from Study 4 with scores from previous literature*

Language	Data reference	SRT (dB SNR)
English	*Study 4 (this study)*	*−11.7 (± 0.3)*
English	Hall, 2006	−10.8
Dutch	Smits et al, 2004	−9.8
Polish	Ozimek et al, 2009	−8.5
German	Wagener et al, 2005	−8.3

Since it is expected that the CRM and the TDT tests are measuring the same hearing abilities it follows that the SRT measured on these two tests should be similar. Table 5.21 shows the difference between the mean SRT for the CRM test conditions in comparison to the TDT. This is calculated by the CRM mean SRT minus the TDT test condition mean SRT (averaged across sessions and repeats), for the normal hearing and hearing impaired samples (a positive value indicates a lower mean SRT for the TDT test condition than the CRM test condition). The CRM-CSoff condition has the closest mean SRTs to the TDT. Overall, individuals found the CRM-CSon condition more difficult than the TDT. This result was to be expected given that in Study 3 the call sign target word group was found to have a higher SRT 50 score than the colour and number target word groups. In order to get the whole sentence correct individuals required a more advantageous SNR, resulting in poorer SRTs for the CRM-CSon condition in comparison to CRM-CSoff.

Table 5.21 *Comparing the SRT scores for normal hearing and hearing impaired listeners on the CRM (CSoff and CSon) and TDT*

Condition	CRM mean SRT minus TDT test condition mean SRT (dB SNR)	
	Normal hearing	Hearing impaired
CRM-CSoff	1.3 (± 0.3)	–0.5 (± 1.0)
CRM-CSon	4.0 (± 0.3)	2.3 (± 0.9)

In summary, the CRM SRTs are correlated with both the TDT and PTA, indicating that the CRM has good concurrent validity. Neither scoring method has displayed significantly better concurrent validity. The scores of the for the CRM-CSoff scoring method are more similar to the TDT than the CRM-CSon method; this indicates that the intelligibility functions of the colour and number target words are more similar to the TDT in comparison to the call signs.

5.5.9 Results: test condition comparisons

Sections 5.5.5-5.5.8 have addressed aims 1-3 of Study 4. The final aim of Study 4 is to decide which of the two CRM test conditions should be further investigated for external validity. This will be assessed in two parts in Section 5.5.9. Firstly, the measurement precision values (stability, repeatability and concurrent validity) for the different CRM test conditions will be compared to assess whether any one test condition 'out performs' another. Secondly, the difference between SRT scores for normal hearing and hearing impaired individuals, as defined by PTA will be investigated to assess whether any one test condition is better at discriminating between the hearing acuity groups, a desirable characteristic of a AFFD measure.

Table 5.22 ties together all the measurement precision results reported in Sections 5.5.5-5.5.8. There is no one CRM test condition which consistently demonstrates better stability, repeatability or concurrent validity. There is therefore no justifiable reason to discard any one test condition for use as a measure of AFFD based on these measurement precision values.

Table 5.22 *Measurement precision summary for each of the CRM test conditions*

Condition	Stability[1]		Variability[2]		Concurrent validity[3&4]	
					With TDT[3]	With PTA[4]
	NH	HI	NH	HI	HI	HI
CRM-CSoff	✓	✓	✓	✓	✓	✓
CRM-CSon	✓	✓	✓	✓	✓	✓

Table notes

1. ***Stability*** Are the mean within session SRT repeats within 1 dB of each other?
2. ***Variability*** Is the 95% confidence interval of the true SRT scores for any one measurement value smaller than that measured for the TDT (NH = < 2.6 dB and HI = < 3.9 dB)?
3. ***Concurrent validity with the TDT*** Is the correlation coefficient (r) 0.65 or higher and is the correlation with the TDT not significantly better in comparisons to the other CRM test condition?
4. ***Concurrent validity with PTA*** Is the correlation coefficient (r) 0.74 or higher and is the correlation with PTA not significantly better in comparisons to other CRM test condition?

Considering that all the CRM test conditions have relatively similar measurement precision properties it follows that if the two test conditions are correlated with each other statistically it would not matter which condition was chosen to be used as a measure of AFFD; the SRT for any one CRM test condition could be used to predict the SRT for any other CRM test condition. In order to assess this, two sets of correlations coefficients have been calculated and are shown in Table 5.23. The correlations in the black shaded cells show the correlation between repeats for each scoring method, the replicability. The correlations in the unshaded cells show the between-test correlations for each repeat. As explained in Section 5.5.1 the replicability of a test is a poor analysis method for assessing measurement precision. However, it is useful value to enable a comparison between how well the CRM test conditions are correlated with themselves and with other test conditions. The correlations have only been carried out for the hearing impaired sample since there is not enough between-subject variation in the normal hearing sample to compute a meaningful correlation (see Section 5.5.1 for details on the limitations of correlation analysis).

The data in Table 5.23 shows that all the CRM test conditions are demonstrating similar levels of correlation for within-test repeat correlations (black shaded cells) and between-test repeat correlations (unshaded cells). In summary, there is no statistical reasoning for choosing one

scoring method over another for further exploration as a measure of AFFD; the measurement precision values are very similar for both of the scoring methods. Furthermore, since the CRM test conditions are as well correlated with between repeats as between condition any one test condition could be selected to predict the SRT scores on the other test condition.

Table 5.23 *Replicability of the CRM for the hearing impaired sample, both within (shaded black) and between (not shaded) test conditions*

Pearson's *r* correlation between repeats and conditions(95% confidence limits lower/upper)				
	CRM-CSoff Repeat 1	**CRM-CSoff Repeat 2**	**CRM-CSon Repeat 1**	**CRM-CSon Repeat 2**
CRM-CSoff Repeat 1		**r= 0.87 (0.71–0.94)**	r= 0.88 (0.72–0.95)	
CRM-CSoff Repeat 2				r= 0.87 (0.71–0.94)
CRM-CSon Repeat 1				**r= 0.91 (0.79–0.96)**

5.5.10 Discussion and conclusions

The overall aim of Study 4 was to assess the measurement precision of the CRM adaptive procedure, with two scoring methods, as a tool for assessing speech recognition in noise. The measurement precision was assessed through evaluating the reliability (including replicability, stability and variability) and concurrent validity of the test; see Figure 5.13 and Section 5.5.1 for an explanation of these terms. Two different scoring methods were explored in Study 4; correctly identifying all three target words (referred to as CRM-CSon) and correctly identifying only the colour and number target words (referred to as CRM-CSoff).

The first part of this discussion will address whether the CRM adaptive procedure test (both scoring methods) has adequate measurement precision, including some discussion about how to define 'adequate'. The second part will include some exploration of how much confidence can be placed on the results, listing the limitations of the study and looking at the agreement between this study and previous research. Finally, the two scoring methods will be compared. Reasons for

choosing one scoring method over another and the potential usage of the CRM in both formats will be discussed.

Measurement precision (reliability and concurrent validity) of the CRM adaptive procedure

There is no definitive method for deciding whether the CRM displays adequate measurement precision, since there are no predefined 'cut off' values for measurement precision of a speech intelligibility test. If the CRM test was 'perfect measurement tool' then it would have been expected that for each subject, when tested under identical test conditions on two different occasions, two equal scores would be recorded. This would result in the stability and variability of the test being 0 dB, which was not the case. Within Study 4 there is no reason to believe there would be a physiological change (e.g. a measurable change in their hearing ability) in the subjects either within or between sessions. A within-session physiological change which impacts an individual's 'true' SRT is unlikely given the short testing period and length of time between repeats. When the NH sample returned for a second session their ears were examined to exclude any visible otological changes and participants were asked if they felt their hearing had changed at all since the previous session; all subject's answered 'no'. This assumption can be verified by comparing the stability and variability values reported in Section 5.5.6 and 5.5.7. For both CRM scoring methods and for the TDT the variability scores (the within-subject standard deviation) is the same size or greater than the stability values. The variability values were all between 1 dB and 3.9 dB whereas the stability values were all < 1dB. This indicates that the change in mean across repeats can be accounted for by random error rather than because of a reproducible systematic effect (Summerfield et al, 1994).

Having established that the results from Study 4 show individual performance to be fairly stable across repeats and that changes between repeats can be accounted for by random error; the next question to ask is whether the amount of random error observed is acceptable when determining SRTs. There is no definitive cut off for 'acceptable' random error but this can be explored by comparing the variability of the two CRM conditions with: 1) the TDT variability results from Study 4 and 2) the variability of other SIN tests in the literature.

The results from Study 4 show both CRM scoring methods as having less variability across repeated measurements than the TDT (shown as the 95% confidence limits of the true SRT score for any one SRT measurement value in Table 5.16). However, the differences between the CRM and the TDT variability values are small (the CRM conditions 95% confidence limits are between 0.5 and 2.2 dB better than the TDT) and as such, this finding should not be used as an indicator that the CRM displays better measurement precision than the TDT. This finding simply shows that the CRM is not showing greater variation between repeats than would be expected for a SIN test.

It is also of interest to compare the variability of the CRM conditions with those of other SIN tests. A review of literature about speech test development has shown that when evaluating SIN tests the focus is placed on measuring whether the test stimuli are of equal intelligibility prior to implementation in an adaptive procedure, rather than measuring repeatability of the adaptive procedure test itself (Kollmeier & Wesselkamp, 1997; Ozimek et al, 2009; Bench et al, 1979). Although the first stage of equalising stimuli intelligibility is an important step in the development of a speech test it should not be assumed that a homogenous speech corpus will guarantee that SRT measured using an adaptive procedure will display good repeatability. Two speech tests which have reported their variability values are the IHR-McCormick Automated Toy Discrimination Test (referred to as Toy Test; Summerfield et al, 1994) and the Dutch TDT (Smits et al, 2004).

Firstly, Summerfield et al (1994) reported the Toy Test to have variability between repeats of 2.5 dB. The Toy Test adaptive procedure was completed by 127 unaided hearing impaired children and the variability value is calculated using the same method as that described in Section 5.5.7. Both CRM conditions have lower variability values than those reported for the Toy Test. Secondly, Smits et al (2004) reported the variability between repeats of the Dutch TDT adaptive procedure by calculating the standard deviation of the difference between repeated measures for each participant. Using this method the Dutch TDT variability value is reported as 1.3 dB when carried out over headphones. The equivalent values have been calculated for Study 4 for both CRM conditions (the within session for the HI sample and both within and between session for the NH sample) so a comparison can be made with the Dutch TDT variability values. Across both CRM scoring methods and for both the NH and HI sample the standard deviation of differences between repeated measures is ≤1 dB. Both the IHR-McCormick Automated Toy Discrimination Test and the Dutch TDT are SIN tests used in clinical environment to assess individuals SIN ability. Considering that the CRM has lower variation in SRT measured between repeats than these two reported tests it is concluded that the CRM has an 'acceptable' level of reliability for use in a clinical environment.

The final topic to review with regards to the measurement precision of the CRM is the concurrent validity of the SIN test when compared to the TDT and PTA. It is expected that if the CRM is in fact measuring SIN ability then there will be a correlation with the TDT, a SIN test, and PTA, a measure of hearing acuity. It has been found that the CRM and TDT not only displayed a good correlation (r = 0.65 and 0.78 for CRM-CSoff and CRM-CSon respectively) but the measured SRTs were also similar across both tests. In addition, the CRM was found to be sensitive to hearing impairment; as hearing acuity decreased (measured by PTA) so did the SRTs. It is therefore possible to conclude that the CRM is measuring SIN ability and is, to some extent, sensitive to hearing loss. These are two key features of a test which could potentially be used to predict military

communication in noisy environments and would be sensitive to discriminating between individuals with normal or impaired hearing.

It is important to reiterate at this point that although the CRM was not found to have markedly better measurement precision than the TDT and the two tests appear well correlated this is not a reason to argue that the TDT could be used as a 'ready to use' alternative. There is no published data about the measurement precision of the TDT and therefore it cannot be viewed as a 'ready to use' SIN test any more than the CRM, despite its widespread use as a screening tool for hearing loss. In addition, one of the key motivations for selecting the CRM over other SIN tests was due to its high face validity for military communications.

To summarise, it can be concluded that, for both scoring methods, the results from this study indicate that the CRM adaptive procedure test demonstrates adequate measurement precision to justify its use to assess speech recognition in noise. This conclusion is based on these three statements:

1) It is expected that majority of variation between repeats can be accounted for by random error and not systematic change, providing confidence in the results.
2) The measurement precision values of the CRM are comparable with the TDT and literature on other SIN test.
3) The concurrent validity displayed between the TDT and the CRM and PTA and the CRM indicates that the CRM is in fact measuring the set of abilities it is designed to (SIN ability and hearing acuity).

Confidence in results and study limitations

The questioning of how much confidence can be placed on the results of Study 4 and considering the limitations of the current study is an important part of deciding what conclusions can be drawn from this experiment. One approach to evaluating this is to compare the results from Study 4 with previous similar studies. This has been reviewed above, with the conclusion that the reliability of the results in Study 4 are not markedly different from those reported in other similar studies (namely Summerfield et al, 1994 and Smits et al, 2004).

One of the limitations of this study was the size of the hearing impaired sample, and the population validity. Population validity refers to whether the sample population represents the population that the results are being generalised to (Shuttleworth, 2009b). This potentially threatens the conclusions that can be drawn from the data. Firstly, due to recruitment difficulties, the hearing impaired population did not contain an even spread of hearing acuities, in particular there were only two participants with severe hearing losses. It is therefore not possible to make

assumptions about how the test performs with this population. Secondly, the data reported about the stability and variability of the CRM for the hearing impaired sample are based on two assumptions: 1) the variation between repeats is equal across the whole hearing impaired sample and 2) the variation between repeats in this sample is representative of the general population. The Bland and Altman Plots in Appendix H did not indicate that the variability between repeats was larger for those with worse SRTs. This may however be a finding that is specific to the small sample explored in Study 4. There is reason to predict that there may be greater variation between repeats for individuals with poorer SRTs. Summerfield et al (1994) measures the word-discrimination thresholds of children and calculated the variability (the within-subject variation) the results across two repeats. They found that as word-discrimination thresholds worsened the variability between repeats also increased. In addition Wilson et al (2007) compared NH and HI adult listeners on the Words-In-Noise test in both stationary speech-spectrum and multi-talker babble noises. They plotted PFs using the data and reported the slopes of the functions for NH listeners were steeper than those for the HI group; this is indicative of HI sample displaying larger variability between responses than those with normal hearing. Due to the limited sample size in Study 4 it is not possible to make an accurate prediction about how variability and stability changes as SIN ability decreases but it is important to acknowledge that the true stability and variability values may be higher than those reported from Study 4 for individuals with worse SRTs.

A second limitation with the study is the lack of between-session data available from the hearing impaired sample; the HI participants only attended one session, completing two repeats within this session. It is therefore not possible to draw any conclusions about the measurement precision of the CRM test conditions for the between-session repeats with the hearing impaired data. However, for the NH sample the stability values between sessions were not markedly different from the within session values. There is no reason to suggest this finding would be different for the hearing impaired sample and it therefore assumed that the measurement precision values calculated from a single within-session repeat would not be significantly different if compared with between-session data.

As aspect of the CRM SIN test which has not been explored in Study 4 is its ability to distinguish between different populations. This is an important issue when considering using the CRM as a measure of AFFD as the test will ultimately be used to discriminate between individuals who are deemed ‘fit for duty’ and those who are not. Ultimately, investigating this is of utmost importance but calculating it with the current data would have required a method for defining the groups for the CRM test to discriminate between. The other method available would have involved using the PTA thresholds to create hearing acuity groups. This was not carried out for two reasons. Firstly, the small sample size and unbalanced numbers of participants with different levels of hearing

acuity would have made it problematic to divide the sample in a meaningful way. Secondly, by evaluating whether the CRM is able to discriminate between different hearing acuity groups would not provide the necessary information about whether the CRM can predict AFFD; in order to do this performance on the CRM needs to be compared to performance carrying out the MCATs listed in Chapter 3. Although it has been established that the CRM is has adequate measurement precision, it is not a useful measure of AFFD if it is not able to accurately discriminate between individuals who can and cannot carry out MCATs to a defined performance level.

At this stage the CRM has only been evaluated in one type of background noise (stationary speech-spectrum) and using one type of measurement procedure (a two-down one-up adaptive procedure). It is possible that future usages of the CRM may involve presenting the CRM stimuli in alternative background noises or using a different measurement method. It is not possible to know whether the CRM would display the same level of measurement precision as reported in Study 4 if aspects of the test were altered. However, it is not possible to measure the measurement precision of the CRM for an infinite number of noise types and measurement methods. If the CRM speech stimuli are implemented in any alternative test formats then an assumption is made that data reported from Study 4 would not significantly vary for the given condition. This may not be a fair assumption for certain types of adaptive procedure or background noises. For example, if an adaptive procedure with larger step sizes was used this may produce a less accurate, and therefore less repeatable, threshold estimate (Kingdom and Prins, 2010).

To summarise, it can be concluded that these initial results show adequate levels of measurement precision for a SIN test but they should be interpreted with caution for two reasons:

1) A larger sample size is needed to predict the stability and variability across a range of hearing losses, giving an indication as to whether these values really are applicable to the entire sample of whether they increase as HL worsens
2) If the CRM stimuli are utilised within an alternative test format to that reported in Study 4 then these measurement precision values cannot be directly applied to other conditions.

Evaluating each scoring method

Two CRM scoring methods were explored in Study 4. The motivation for this was two-fold. Firstly, introducing and validating two scoring methods would broaden the potential future uses of the CRM. Secondly, since the call sign target word group was found to be the least homogeneous (see Study 3) there was some concern that scoring responses to the call sign target word may reduce the overall measurement precision of the CRM adaptive procedure. The results in Study 4 have

not shown either scoring method to display significantly better measurement precision than another (see Table 5.22). It was also shown that the two scoring methods are well correlated with each other, indicating that one test condition could be used to predict performance on another. It is therefore not necessary to reject either scoring method based on poor measurement precision. It can be concluded that the call sign target word group was 'homogenous enough' for the reliability of the CRM adaptive procedure to remain unaffected. This finding extends the options for how the CRM test is used.

The results of Study 4 do not provide a definitive answer about which scoring method should be further explored as a measure of AFFD. Inclusion of the call sign arguably gives the test higher face validity since in an operational scenario individuals are required to hear all parts of a command. However, since the addition of the call sign does not improve measurement precision the test could be run in its simplest format, CRM-CSoff; the addition of the call sign does not provide any specific further information about the individual's SIN ability and slightly increases the testing time. Ultimately, the CRM test condition which is best able to predict performance on MCATs should be selected as a measure of AFFD and this information is not available from the results of Study 4; Chapter 6 provides further information about this.

Conclusions

To summarise, two concluding statements can be drawn from Study 4. Firstly, both scoring methods (call sign on and off) for the CRM adaptive procedure test in stationary speech-spectrum noise have adequate measurement precision to be used to measure individual SRTs. Secondly, the suitability of the CRM as a tool for assessing AFFD remains unknown. Further work is required to explore the relationship between performance on the CRM and performance carrying out the SC-MCATs.

5.6 Chapter 5 Summary

Chapter 5 has reported the entire process of selecting, developing, recording and evaluating the measurement precision of a SIN test. At this stage it has been concluded that the CRM adaptive procedure in stationary speech-spectrum noise is a 'ready to use' SIN test which displays good reliability and concurrent validity with two scoring methods. In addition, it has been shown that the CRM is sensitive to hearing impairment. On average as an individual's pure-tone thresholds increase their score on the CRM decreases. At this stage it is still not possible to state whether the CRM is a suitable tool for predicting AFFD. There is no evidence which links performance on the CRM to performance on the SC-MCATs. In order for the CRM to be considered for use as a measure of AFFD it is necessary to explore the association between scores on the CRM and

performance when carrying out the SC-MCATs identified in Chapter 3. Steps towards achieving this are addressed in Chapter 6.

Chapter 6: Developing a method to assess performance on the speech communication MCATs

6.1 Introduction

Chapter 5 has explained the development of the CRM and that it is now considered to be a 'ready to use' SIN test. The next step towards utilising this test as a measure of AFFD is to assess the predictive validity of the CRM as measure of AFFD. There is currently no evidence looking at the relationship between performance levels on the current measure of AFFD, PTA, and performance on the MCATs. The first stage towards addressing this knowledge gap is to develop a method(s) for measuring performance on the MCATs and this is the focus of Chapter 6. Since the focus of this thesis is on the SC-MCATs, Chapter 6 will explore methods for assessing performance on the SC-MCATs (see Table 6.1 for a recap of these), with the ultimate view of using this assessment tool(s) for measuring and comparing the predictive validity of the CRM and PTA as measures of AFFD. It is worth remembering that the SC-MCATs identified in Chapter 3 are those carried out by infantry and combat-support personnel and therefore…

Table 6.1 *Speech communication MCATs*

T1	Accurately hearing commands in a casualty situation
T2	Accurately hearing grid references
T3	Accurately hearing directions on patrol
T4	Accurately hearing directions in a vehicle
T5	Accurately hearing fire control orders
T6	Accurately hearing 'stop' commands
T7	Accurately hearing the briefing before a foot patrol

The chosen method for measuring the predictive validity of AFFD assessment tools must have high external validity; that is to say that the results from the test must be generalisable to performance in real world operational environments. Figure 6.1 explains how using an externally valid test will allow for predictions to be made about the relationship between the performance when carrying out the SC-MCATs in a real world operational environment and the performance on the CRM and PTA.

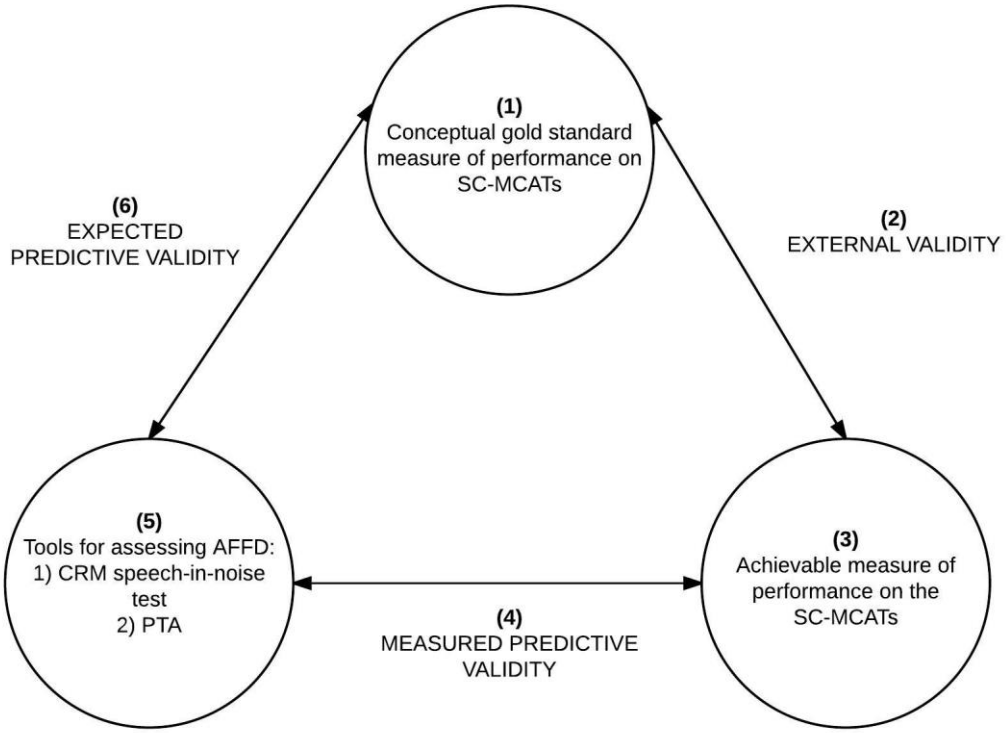


Figure 6.1 *Diagram showing the relationship between a gold standard measure and an achievable measure of performance on the speech communication MCATs (SC-MCATs) and how a clinical speech-in-noise test may be used to predict performance on either of these. Numbers relate to description in text directly below figure.*

The gold standard measure of performance on the SC-MCATs would involve assessing individuals carrying out each different task, across a range of scenarios in a real world operational environment (Figure 6.1, point 1). However, running an experiment in this environment is not possible, and it is therefore necessary to consider alternative achievable testing options (Figure 6.1, point 3) which hold high external validity to performance in real world settings (Figure 6.1, point 2). If scores on the CRM or PTA (Figure 6.1, point 5) are shown to be able to accurately predict performance carrying out the SC-MCATS (Figure 6.1, point 4) and the SC-MCATs performance test has high external validity (Figure 6.1, point 2), then it is expected that the CRM or PTA will to some extent be able to predict performance whilst carrying out the SC-MCATs in a real world listening situation (Figure 6.1, point 6).

External validity of a test relates to the generalisability of the test results; to what extent can the results be generalised to other populations or settings (Campbell & Stanley, 1966). External validity is usually split into two distinct types: 1) ecological validity and 2) population validity (Shuttleworth, 2009a). Appendix A provides an overview of all the definitions of all types of validity referred to in this thesis. Ecological validity refers to the degree to which observations recorded within a study reflect performance that would occur in a real world setting; it is this type that is of most interest when selecting a measurement method to assess performance on SC-

MCATs. Section 6.2 reviews a variety of methods for measuring performance on the SC-MCATs, considering the ecological validity of each. Population validity evaluates whether the sample population represents the population that the results are being generalised to (Shuttleworth, 2009b). When developing a test to assess performance on the SC-MCATs it is important that the test is evaluated with a sample of military personnel since it is predicted that performance carrying out the SC-MCATs will be affected, to some extent, by factors such as job experience or knowledge of military specific language or vocabulary.

Chapter 6 explores the various methods for measuring performance on the SC-MCATs, leading to the justification for the chosen method (Section 6.2). Study 5 (Section 6.3) details the design and development of the chosen method, the Vehicle Communication Simulated MCAT (VEHCOM SimMCAT). Study 6 (Section 6.4) reports the initial assessment of the VEHCOM SimMCAT as a tool for predicting performance on the SC-MCATs, specifically focusing on whether it is sensitive to hearing impairment and job experience, factors that are expected to influence performance on the SC-MCATs. Towards the end of Section 6.4 the steps required to move towards using the VEHCOM SimMCAT as a method for assessing the predictive validity of the CRM and PTA as measures of AFFD, the ultimate aim of this thesis, are explored. Finally, Section 6.5 provides a summary of this chapter.

6.2 Methods for measuring performance on the SC-MCATs

In order to achieve the ultimate goal of measuring and comparing the predictive validity of the CRM and PTA as measures of AFFD, a method for assessing performance on the SC-MCATs is required. In Section 6.2 various methods for measuring performance on these tasks will be explored. The selected method must have high external validity in relation to real world operational scenarios. An aspect of this is population validity and as such any assessment method should be evaluated with military personnel with varying levels of hearing acuity. It is pre-empted that recruiting a large enough sample of personnel to represent a wide enough range of hearing acuities to conduct this evaluation would be problematic, based on experience from running Study 4 (see Appendix O for details of recruitment challenges in the military population). Using a sample taken from the hearing impaired civilian population to make predictions about AFFD performance within a military population would result in low population validity. Using hearing loss simulation (HLS) allows for a sample of normal hearing military personnel (who are easier to recruit) to be tested and the results used to predict performance levels of hearing impaired listeners. For this reason, the scope for incorporating HLS technology into potential methods for measuring performance on the SC-MCATs; this is explored in Section 6.2. The potential methods can be split into three categories:

1. real world training scenarios
2. task simulation using military simulation equipment
3. task simulation set up in a clinical environment

These are further explored in sections 6.2.1 to 6.2.4, detailing the basic premise of each method, the potential scoring methods, and the advantages and disadvantages to the method.

6.2.1 Live training scenarios

Explanation: A set of real world training exercises which are based on the tasks listed in Table 6.1 and are closely related to training exercises are already used in practice. For example, for T4, accurately hearing directions in a vehicle, an exercise could involve personnel completing a specified driving route whilst being instructed over radio which direction to move. Or, for T2, accurately hearing grid references may require personnel to move to a specified location or mark it on a map. This method would arguably have the highest external validity, in comparison to the other two methods discussed in Sections 6.2.2 and 6.2.3.

Example: Combat effectiveness with varying levels of hearing acuity was measured using a paintball-based simulated military exercise by Brungart and his team at the Walter Reed National Military Medical Center, Washington, USA (Brungart et al, 2013). Participants included normal hearing military personnel who wore real-time electronic HLS systems whilst conducting a militarily-relevant task with a main aim to eliminate opposing players (lethality) and to be the last player remaining (survivability).

Scoring methods: For this type of method a scoring system is required which can measure the performance for each individual. Brungart and his team (Brungart et al, 2013) scored individuals on how many opponents they eliminated and how long they lasted in the game (based on the order of elimination). Discussions with subject-matter experts who are involved in the design and running of training exercises at the Infantry Trials and Development Unit (ITDU), Warminster, revealed that there are no quantifiable scoring methods used to assess individual performance on training exercises. The majority of training exercises are completed as part of a mission, and therefore the overall outcome of the exercise is focused around mission success, rather than individual performance. Investigation was carried out to determine the methods used to assess hearing protection devices, since it was thought that perhaps there would be a validated method for assessing how well individuals performed when wearing different devices, but it was found that the methods were subjective and not quantifiable.

Advantages: Since military training exercises are designed to prepare individuals for real world operational scenarios, this method would provide the closest predictions of personnel's performance when they are deployed. If the 'gold standard' measure of performance on the SC-MCATs is to assess performance in an operational setting, then arguably measuring performance through training scenarios which are closely linked to experiences during deployment is the closest feasible alternative.

Disadvantages: It would be difficult to control the environment in which testing was carried out; making comparisons between individuals/groups may be problematic. For a number of training exercises levels of experience may greatly impact performance, which may make it difficult to make comparisons between groups and/or individuals. Hearing loss simulation would need to be conducted in real-time. Real-time wearable HLSs are not readily available in the UK, meaning a normal hearing sample could not be recruited as an alternative to hearing impaired personnel. There are no validated quantifiable scoring methods which could be readily used to measure individual performance in operational settings.

6.2.2 Simulated training scenarios

Explanation: The Armed Forces use simulation equipment to train personnel, allowing personnel to train for mission success without the impracticalities and costs of live training. Simulation equipment could be used to simulate some of the MCATs listed in Table 6.1, allowing for performance on these tasks to be measured in a more controlled environment than using live training scenarios.

Examples: The two main simulation facilities used by infantry and combat support personnel are the Dismounted Close Combat Trainer (DCCT) and Combined Arms Tactical Trainer (CATT), which are setup at ITDU. The DCCT (Figure 6.2) is a computer game type setup on a large projector screen where groups are tested on their team work and can complete live-fire training (Meggitt, 2015). It is possible to envisage using this sort of equipment to measure performance on T5, for example (accurately hearing fire control orders). The CATT (Figure 6.3) is a collection of simulated vehicles and control stations which have replica interiors to real vehicles, such as tanks and helicopters. This equipment is used to test team work skills, communication and instruction following ability, as well as interaction with civilian/friendly forces (Lockheed Martin, 2015). The CATT could potentially be used to replicate scenarios relating to T2 (accurately hearing grid references), T4 (accurately hearing directions in a vehicle) and T6 (accurately hearing 'stop' commands). With both the DCCT and CATT it may be possible to pre-process stimuli through a HLS, eliminating the need for real-time HLS technology.

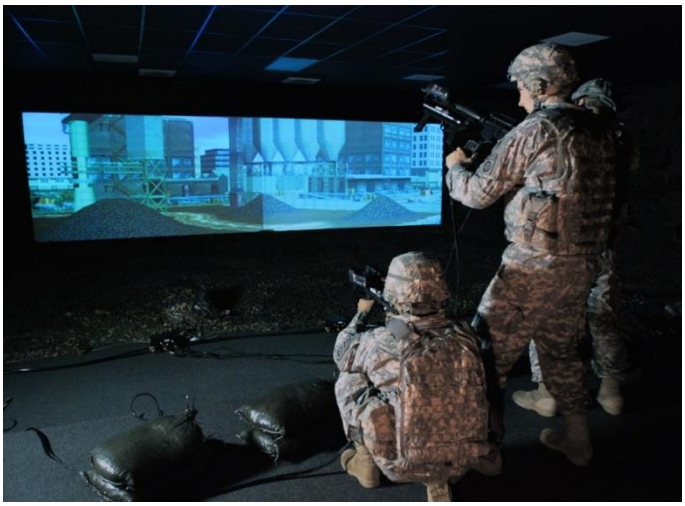


Figure 6.2 *Dismounted Close Combat Trainer (DCCT) (Espaillat and Smith, 2010)*

Figure 6.3 *Combined Arms Tactical Trainer (CATT), inside a tank simulator (Wallace, 2012)*

Scoring methods: The same problems relating to scoring methods discussed in Section 6.2.1 are encountered when using simulated equipment. Discussions with the individuals who run the simulation equipment revealed that performance is measured across a team, not for individuals. However, in this more controlled environment it is possible to envisage designing a scenario in which individuals could be quantitatively scored on how well they perform a set task. For example, did they follow the correct directions whilst driving the vehicle or did they carry out the correct fire control order.

Advantages: The use of simulation equipment allows for more scope to control the environment in which testing occurs, in comparison to live training scenarios. This means that large numbers of participants could complete identical test procedures; this would mean comparisons between individuals/hearing acuity groups would be possible. There is potential for using HLS technology as speech signals could be pre-processed, preventing the need for real-time equipment.

Disadvantages: This equipment has been designed to replicate specific military training scenarios and as such there is limited scope for creating new scenarios relating to the SC-MCATs. There are no quantifiable scoring methods routinely used with the equipment; feedback is normally subjective and provided to an entire group, rather than individuals. This method is a step away from measuring performance in a 'real world' operational setting, and as such assumptions must be made about the relationship between performance on simulated training equipment and performance in a combat scenario. On a practical note, the facilities at Warminster are used almost continuously for military training and as such it is unlikely the equipment would be available for experimental purposes.

6.2.3 Task simulation in a clinical environment

Explanation: Scenarios which relate to the SC-MCATs could be created and run in a controlled clinical environment, without the use of specialist military equipment. This would involve listening to commands which relate to the SC-MCATs and responding in a measurable format. Since personnel are rarely only completing one task at a time a second visual response task could also be introduced.

Examples: Participants would be required to listen to commands which relate to the SC-MCATs and respond by either repeating what they have heard or responding on a computer screen. The simplest version of this task would involve being scored on accuracy of command repetition. A more complex option could involve responding to the command with an appropriate action. For example, they may be presented with a direction and asked to select an arrow which depicts the instructed direction of travel, or they may be presented with a grid reference and asked to select the corresponding area on a map or type the numbers. An additional task, competing for the individual's attention, could be introduced, such as a visual awareness task with varying levels of difficulty. An example visual awareness task has been developed and is detailed in Appendix I. The presented commands could be pre-processed using HLS.

Scoring methods: Individuals could be scored on the percentage of correctly repeated words or appropriate responses to commands. Separate scores could be recorded depending on whether a second visual awareness task was being completed at the same time and the difficulty level of the additional task.

Advantages: Compared to the previous two methodologies (sections 6.2.1 and 6.2.2) this would be the simplest experiment to run in terms of set-up and recruitment since it is not dependent on coordinating availability of training grounds/simulation equipment and personnel's diaries. It also allows provides the researcher with complete autonomy over the testing conditions, since there are no constraints with regards to equipment limitations or possible training scenarios. It is also easy to control the testing procedure, ensuring that all participants undergo identical testing conditions, allowing for comparisons to be made between hearing acuity groups.

Disadvantages: This testing method is the furthest removed from a real world operational environment and as such assumptions must be made about the relationship between performance on this method and performance in a combat scenario. Without validating the tool it is not possible to draw any definitive conclusions about the impact of hearing impairment on operational performance. Validation of the tool would be a lengthy process and would require a 'gold standard' measure of military communication, which is not available.

6.2.4 Chosen method: justification and limitations

Sections 6.2.1-6.3.3 have introduced three potential methods which could be used to measure performance on the SC-MCATs and as a tool to assess the predictive validity of the CRM and PTA as measures of AFFD. Figure 6.4 shows the compromise to be made when selecting a measure of military communication performance. The performance measures which have the strongest association with the real world job are also the measures which would be the most difficult to develop, control and run, and vice-versa.

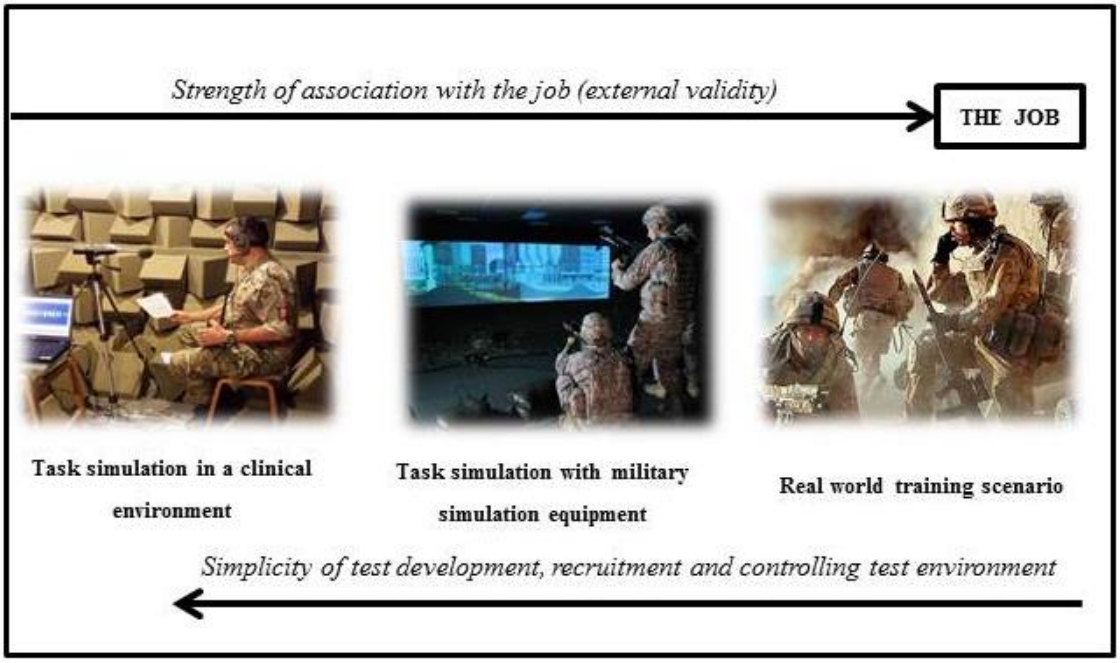


Figure 6.4 *The compromise to be made when selecting a measure of military communication performance between a test with a high ecological validity and a test which is simple to develop, run and control.*

Rather than selecting a measure of military communication with the highest external validity that would be difficult to develop and run, the simplest testing method will be developed, which can then ultimately be used to provide an initial indication about the relationship between performance on the CRM and PTA, and performance in real world listening scenarios. Simulated tasks which relate to the SC-MCATs will be developed and run in a clinical environment, without using any specialist military equipment. Although it is understood that personnel rarely only carry out a single task at any one time, at this early stage the focus will primarily be on replicating the auditory elements of the SC-MCATs alone. Developing a test that represents the auditory elements of the SC-MCATs will provide a baseline tool that could be advanced to become a more comprehensive measure with higher external validity to the task.

It is important to distinguish at this point the difference between the *task* and *environment* elements of the SC-MCATs, since this influences the design and applicability of any chosen method. The SC-MCATs are a list of specific speech-communication *tasks* that can be completed in a variety of *environments*. Multiple SC-MCATs could be completed in the same environment and in conjunction with each other. For example, in the environment of a combat scenario with weapon noise, personnel could realistically be carrying out multiple SC-MCATs in conjunction with one another, such as T1 (accurately hearing commands in a casualty situation), T5 (accurately hearing fire control orders) and T6 (accurately hearing 'stop' commands). When considering developing a task simulation to measure performance on the SC-MCATs there are three options. Firstly, develop multiple simulated environments which allow performance on all the SC-MCATs to be measured. Secondly, develop a single simulated environment which relates to one SC-MCAT. Thirdly, develop a single simulated environment in which a number of the SC-MCATs would be carried out. It has been decided that for the initial development of a tool for assessing performance on the SC-MCATs the third approach will be used, since this will create a tool that will allow for a general assessment of performance on multiple SC-MCATs, albeit in only one environment.

It is acknowledged that there are limitations to the chosen methodology; it is important to recognise that it is not being claimed that it will provide all the information required to predict how well an individual would perform on all the SC-MCATs. The aim is to design and develop a tool that can ultimately be used to give an initial indication about the relationship between the CRM and PTA, and performance on tasks which relate to the SC-MCATs. However, there are limitations to the chosen methodology, which are broken down into three categories:

1. ***The methodology is only assessing auditory component of SC-MCATs***

All of the SC-MCATs identified in Chapter 3 require personnel to do more than just listen to auditory commands. At the same time as listening to the auditory commands, personnel are often carrying out additional tasks such as driving a vehicle, reading a map or acting on fire control orders, whilst at the same time maintaining situational awareness. It is known that when an individual carries out two or more tasks at the same time, attention capacity limitations influence their performance levels for each task (e.g. Fulcher, 2003). An example of this is outlined by Strayer and Johnston (2001), who carried out a dual-task study assessing the effects of holding a conversation over a mobile-phone on performance on a simulated driving task. They found that failure to detect simulated traffic signals and reaction times to detect these signals was increased two-fold when individuals were simultaneously holding conversations using either a handheld or hands-free phone. It is assumed that a similar deterioration in performance would be observed

when personnel have to carry out more than one task at once for the SC-MCATs, but this will not be taken into account using the proposed method of only measuring performance on the auditory element of the SC-MCATs.

2. The test is only assessing performance in a single environment

Developing a measure of performance on the SC-MCATs that only focusses on a single environment will provide a general idea about performance on multiple SC-MCATs, but the applicability of the results will be limited to the specific environment and the SC-MCATs that are carried out within it.

3. The external validity of the test will remain unquantified

The association between performance on the test proposed in Section 6.3 and performance carrying out the SC-MCATs in a real world operational setting will not be measured and will therefore remain unquantified; assessing 'real world' performance on the SC-MCATs (the right side of Figure 6.4) is problematic. This raises an issue when it comes to making suggestions for 'cut-off points' determining whether an individual is fit for duty or not.

6.2.5 Summary

Section 6.2 has investigated the variety of possible methods for measuring performance on the SC-MCATs. The chosen method will involve developing a task simulation which can be run in a clinical environment. This will involve developing a single simulated operational environment in which a number of the SC-MCATs would realistically be carried out. Although there are limitations to this method (see Section 6.2.4), it has been decided that this method will provide a compromise between developing a test which is able to provide a good initial indication of performance on the SC-MCATs, but without being overly difficult to develop and run (see Figure 6.4). Section 6.3 will now explain the methodology for developing the SC-MCAT simulation test.

6.3 Study 5: Developing, recording and evaluating the Vehicle Communication Simulated MCAT (VEHCOM-SimMCAT) test

6.3.1 Introduction

In Section 6.2 the chosen methodology for measuring performance on the SC-MCATs has been justified. A test will be designed that can be carried out without any specialist equipment and will involve listening and responding to commands which relate to the SC-MCATs. Section 6.2.4 established that a task simulation that focusses on a specific environment in which multiple MCATs would be carried out will be developed. In order to decide which environment the simulation should be focused, on the author met with a subject-matter expert at ITDU (who had operational experience) to determine an operational scenario in which the majority of the SC-MCATs would be carried out. It was decided that when listening to commands over a radio in a moving vehicle T1, T2, T4, T5 & T6 would realistically be conducted. This scenario was deemed simple to control acoustically since the background noise (a vehicle engine) is fairly continuous, making it easier to control the SNR of the test. The design and development of a test that measures performance on the SC-MCATs carried out when communicating over radio in a moving vehicle is detailed in Section 6.3.3. The test is named the Vehicle Communication Simulated MCAT test, which is abbreviated to VEHCOM SimMCAT for the remainder of this thesis. Section 6.3.4 contains an explanation of the finalised test format.

6.3.2 Research objective 5

There is no 'knowledge gap' as such which relates to this research objective. The motivation behind Study 5 is that in order to measure the predictive validity of the CRM and PTA, as measures of AFFD, a test is required which is able to measure individual performance on the SC-MCATs. Study 5 will address the design and development of a test that is able to achieve this, specifically focusing on the scenario of listening to commands over a radio in a moving vehicle.

Research objective 5: To design and develop a simulation for measuring performance on the auditory element of the SC-MCATs, focusing on the scenario of listening to commands over a radio in a moving vehicle.

6.3.3 VEHCOM SimMCAT development

Section 6.3.3 explains the stages of the VEHCOM SimMCAT development; an overview of the process is given in the flow diagram (Figure 6.5). The remainder of this section is numbered in accordance with the stages outlined in the flow diagram.

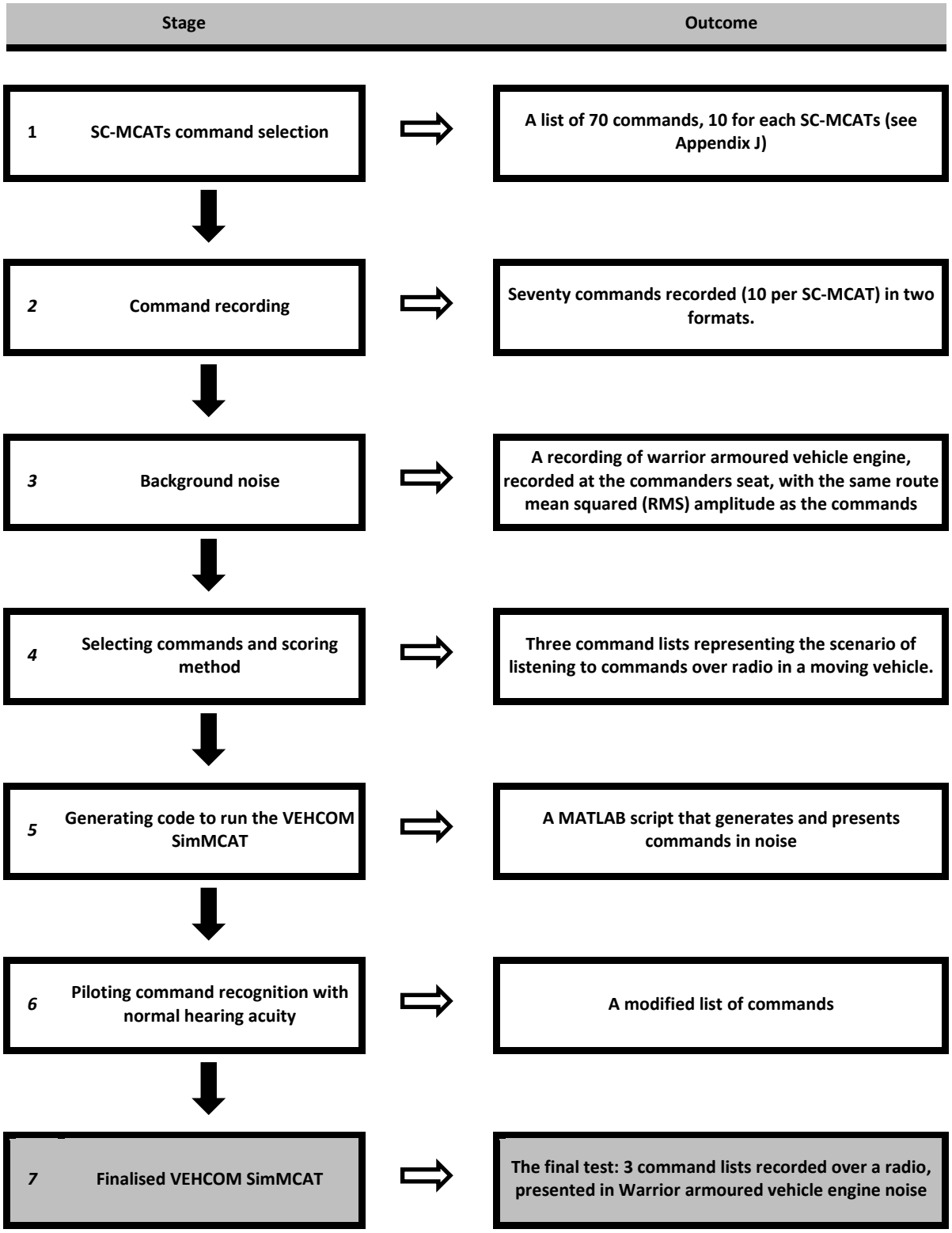


Figure 6.5 *Flowchart showing the development stages for the VEHCOM SimMCAT*

1. Speech communication MCATs command selection

Although not all of the SC-MCATs will be assessed in the VEHCOM SimMCAT it was decided that this opportunity should be used to create test material for all of the SC-MCATs, increasing scope for future research. In order to measure performance on the auditory aspect of the seven SC-MCATs identified in Chapter 3 (see Table 6.1) a list of commands which related to these tasks was required. These were developed in correspondence with SMEs from ITDU, Warminster.

1.1. *Command list development with subject matter experts (SMEs)*

A Warrant Officer from ITDU was recruited to formulate a list of ten commands that related to each SC-MCAT. It was decided that ten commands for each SC-MCAT would be a sufficient number to allow for enough variety but without creating a speech corpus which was too large and difficult to manage.

1.2. *Verification of command list with additional SMEs*

This list was forwarded via email to the Medical Squadron from the Commando Logistic Regiment (Barnstaple) who was asked to confirm that the commands and their format was generic across different regiments. Some minor amendments to the original format of the commands were made and some additional commands were added to some tasks.

1.3. *OUTCOME: a list of commands for each SC-MCAT*

A final list of commands for each SC-MCAT was generated (see Appendix J) to be taken forward to the recording stage.

2. Command recording

The commands were spoken by the same individual who developed the command list. The Warrant Officer had extensive experience issuing commands to infantry and combat-support personnel both in a training environment and during deployment and therefore had an understanding about how to articulate the words and intonate the sentences. The commands recorded are those listed in Appendix J.

2.1. *Recording the commands*

The commands were recorded at the University of Southampton using the set up shown in Figures 6.6-6.8). The recording setup was such that each command was recorded in two formats at the same time: 1) live voice and 2) through a Personal Role Radio (PRR) headset. Three repeat recordings were made of each command. Although only the recordings through the PRR would be required for the VEHCOM SimMCAT, it was logical to use this opportunity to make live voice recordings at the same time, increasing future potential uses.

For the live voice recording the speaker was sat in an anechoic chamber 0.75m from a Brüel and Kjær Precision Integrated sound level meter (SLM), type 2230, which was recording his live voice (see Figure 6.6). The SLM was connected via an RME Babyface 22-Channel soundcard to a Windows XP Laptop, running Adobe Audition (v3.0) which was used to record the speech. The commands were recorded as mono sound files at a 44100 Hz sampling rate.

For the recordings through the PRR the speaker was wearing a PRR which was being transmitted through the wall of the anechoic chamber to a separate listening room, where a Knowles Electronic Manikin for Acoustic Research (KEMAR) was situated wearing a second PRR (see Figure 6.13). The spatial separation between KEMAR and the speaker was necessary to ensure no live voice was recorded by KEMAR. KEMAR wore the PRR headset on his right ear (not as photographed in Figure 6.8, the photograph was taken at a separate time to recording). The G.R.A.S. KEMAR Head and Torso simulator was fitted with G.R.A.S. IEC 711 RA0045 ear simulators (including 40AG ½" microphones). These were connected via a G.R.A.S. 12AK 1-channel power module and RME Babyface 22-Channel soundcard to an Apple Mac computer running Adobe Audition (v3.0) which was used to create the recordings. The commands were recorded as mono sound files at a 44100 sampling rate. The KEMAR manikin allowed for recording the PRR in-situ. KEMAR simulates the changes that occur to sound waves as they pass a human head and torso, allowing for recordings of the sound that reaches the ear drum when personnel are wearing PRR headsets.

Three recordings of each command were made allowing for the best version to be selected at the end of recording. Commands were removed if the speech was not pronounced clearly and/or the commands were said unrealistically quickly and/or the recording contained interruptions, such as a coughs or rustling.

2.3. *Equalising the RMS amplitude of all the commands*

MATLAB (R2013b) was used to equalise the RMS amplitude of all the command recordings (separately for the clean speech and PRR recordings). The same script as used in Appendix E was used to achieve this.

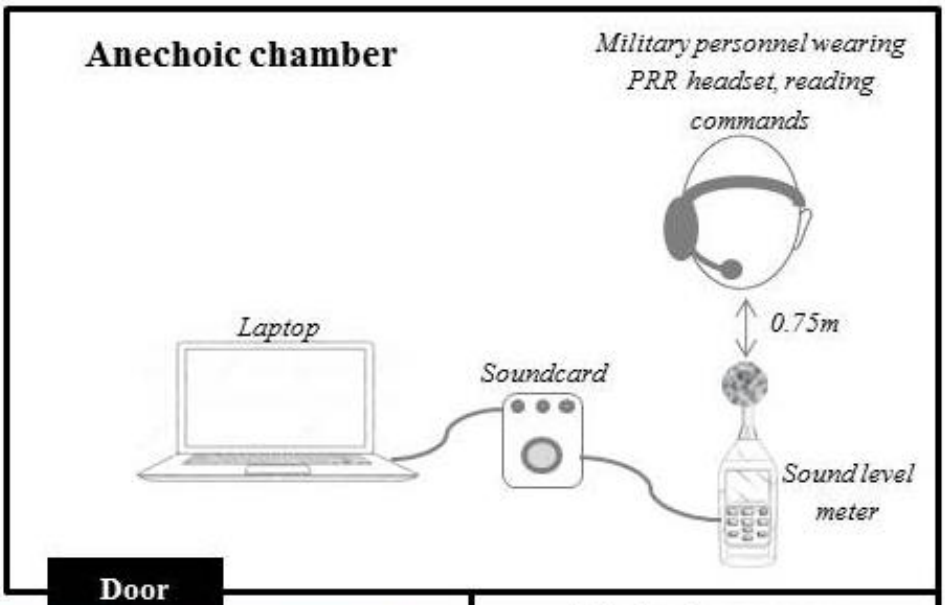
2.4. *OUTCOME: recorded commands in two formats, clean speech and over a PRR.*

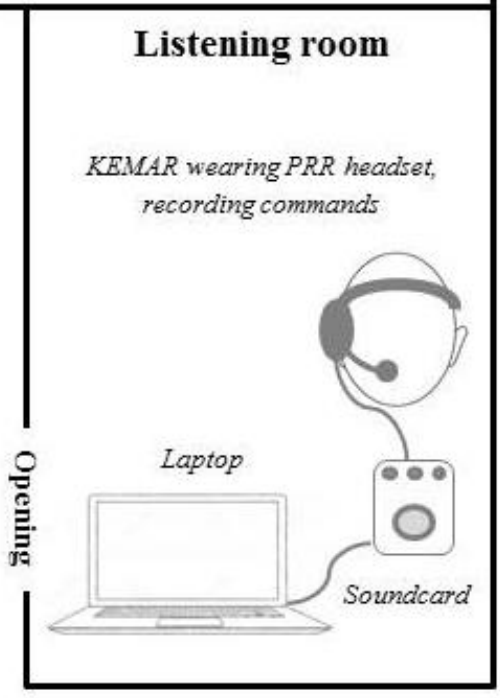
Each recording was saved as an audio file ready for use.

Figure 6.6 *(right) Recording set-up. Speaker in anechoic chamber*

Figure 6.7 *(below) Recording set-up. Diagram of room layout and equipment*

Figure 6.8 *(bottom left) Recording set-up. KEMAR set-up in listening room*





3. Background noise selection

The remainder of Section 6.3.3 will focus specifically on developing the VEHCOM SimMCAT. An appropriate vehicle engine background noise for this simulation needed to be selected, obtained and processed ready for use in the test. The chosen vehicle needed to be one in which personnel communicate over a radio headset and one that the majority of infantry and combat-support personnel would travel in (to avoid developing an VEHCOM SimMCAT that is not generalisable). The chosen vehicle was a Warrior armoured vehicle; an infantry fighting vehicle currently in service with armoured infantry battalions (Defence Suppliers Directory, n.d.). The armoured vehicle accommodates three crew and seven dismounted infantry personnel, and when riding inside the vehicle personnel communicate over radio headsets (British Army, n.d.).

3.1. *Obtaining recordings and equalising the RMS amplitude*

Noise recordings from a standard FV510 Warrior Armoured vehicle were obtained from a library of vehicle engine noises available at the Institute of Naval Medicine Acoustics Department. The recordings were made whilst the vehicle was driven over road at Salisbury Plan, near Warminster at 20 km/hour. Although recordings were available from a variety of surfaces and different positions within the armoured vehicle (see Figure 6.9), the recording of the vehicle being driven on a road (tarmac and concrete) was selected as this contained the least fluctuations in level caused by changes in terrain. The recordings from the commander position were selected as this person will be required to receive and issue commands whilst in the vehicle. A 15 minute noise recordings was made at the approximate ear point (about 15 cm from the left ear) at the commander position in the Warrior armoured vehicle (see Figure 6.9), using a 1/2'' condenser microphone (G.R.A.S. type 40AR).

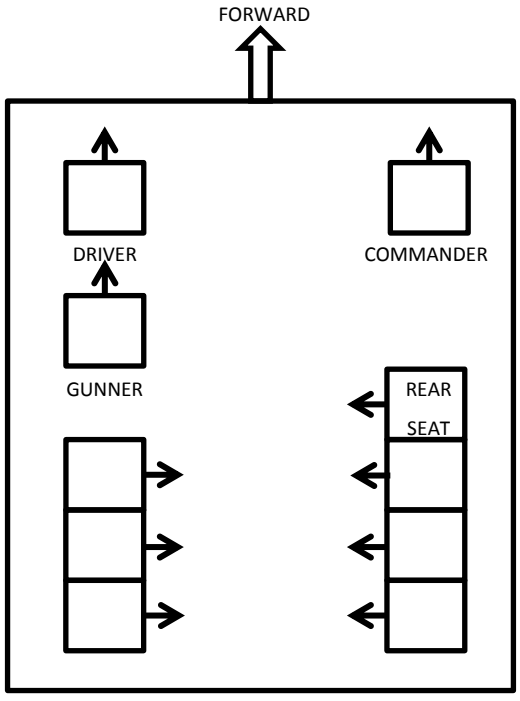


Figure 6.9 *A schematic diagram of the positions within Warrior armoured vehicle, showing the commander position where the microphone was placed to record the engine noise.*

3.2. *OUTCOME: Warrior engine noise to be used in the VEHCOM SimMCAT*

A 15 minute long audio recording of engine noise recorded at the position of the commander is available for use in the VEHCOM SimMCAT. The recording was made in a standard FV510 Warrior Armoured vehicle being driven at 20km per hour over road and taken from the commander position (15 cm from the left ear). The RMS amplitude of the noise was equalised to be the same as the commands (see Appendix E).

4. Selecting commands for the VEHCOM SimMCAT and selecting a scoring method

Not all of the SC-MCATs would be conducted over a PRR or in the presence of engine noise. A selection of commands needs to be chosen which would realistically be issued in this scenario.

4.1. *Selecting commands which would be issued in a vehicle over a radio headset*

The first stage of selecting commands appropriate for the chosen scenario involved deciding which SC-MCATs could feasibly be conducted in the presence of vehicle engine noise and over a PRR. After a discussion with members of the Hear for Duty Team at Southampton it was decided that five of the seven SC-MCATs could feasibly be conducted in this scenario:

- (T1) Accurately hearing commands in a casualty situation
- (T2) Accurately hearing grid references
- (T4) Accurately hearing directions in a vehicle
- (T5) Accurately hearing fire control orders
- (T6) Accurately hearing stop commands

The second stage was choosing appropriate commands to include in the VEHCOM SimMCAT. The author selected a number of commands from each MCAT, focusing on selecting commands which are most likely to be spoken whilst in a vehicle and over a radio headset. In addition, from each SC-MCAT, commands which were similar length were selected. A subject-matter expert from ITDU (also involved in developing the command list and recording the commands) confirmed over email that all of the selected commands would realistically be issued over radio whilst in a vehicle. The selected commands are highlighted in Appendix J.

4.2. *Considering different scoring methods*

In order to measure performance on the VEHCOM SimMCAT a scoring method was required. The options for measuring performance could be split into two categories: 1) comprehension/understanding of the command and 2) repetition of the command.

Command comprehension is arguably the most externally valid of the two options; in a real world scenario personnel are required to understand and appropriately act upon commands. The options for measuring this would involve recording a response to the command issued. This could be simple for commands such as “at the fork in the track go left”; one can envisage a graphical user interface on which the individual is required to select which way to turn. However for

commands such as “we have three casualties” or “enemy forces seen at edge of building”, which do not require an obvious response this method would prove difficult. In addition it may prove difficult to measure the difference between whether an individual did not hear the command or was not sure how to respond appropriately (because of lack of experience or the response task being unclear); this would make interpreting the results difficult.

Although it is understood that key word repetition (measuring percentage correct) is not how personnel would respond to the commands in a real world listening scenario it was decided that this would be the simplest performance measure at this early stage of VEHCOM SimMCAT development. This method is commonly used for measuring speech intelligibility in noise in a clinical environment.

Key words and phrases were assigned to each command and participants were scored on whether they correctly repeated a key word/phrase. Within each command any word which was either providing information or if it was removed would mean the command could not be understood was classified as a key word/phrase. In some instances it made more sense to group together two or three words as a key phrase than to score each of those words individually. One point was available for a key phrase such as, ‘go left’, ‘my location’, ‘gunshot wound’, meet liaison officer’ and ‘await my order’. One point was available for each grid reference, requiring all six digits to be correctly repeated; in a real world scenario only hearing part of a grid reference is not sufficient to accurately carry out a task. It is acknowledged that there is no definitive method for selecting key words/phrases and the final choices were confirmed with other members of the Hear for Duty team. In Appendix J the key words/phrases are shown by bold text, separated by slashes for each of the highlighted commands.

5. Generating the VEHCOM SimMCAT test

The next stage of test development involves taking the audio files containing the recorded commands and the vehicle engine noise audio file and combining these to develop a test which can be used to score performance (percentage of correctly repeated key words/phrases) when listening to the commands at a specified SNR.

5.1. *Creating three command lists*

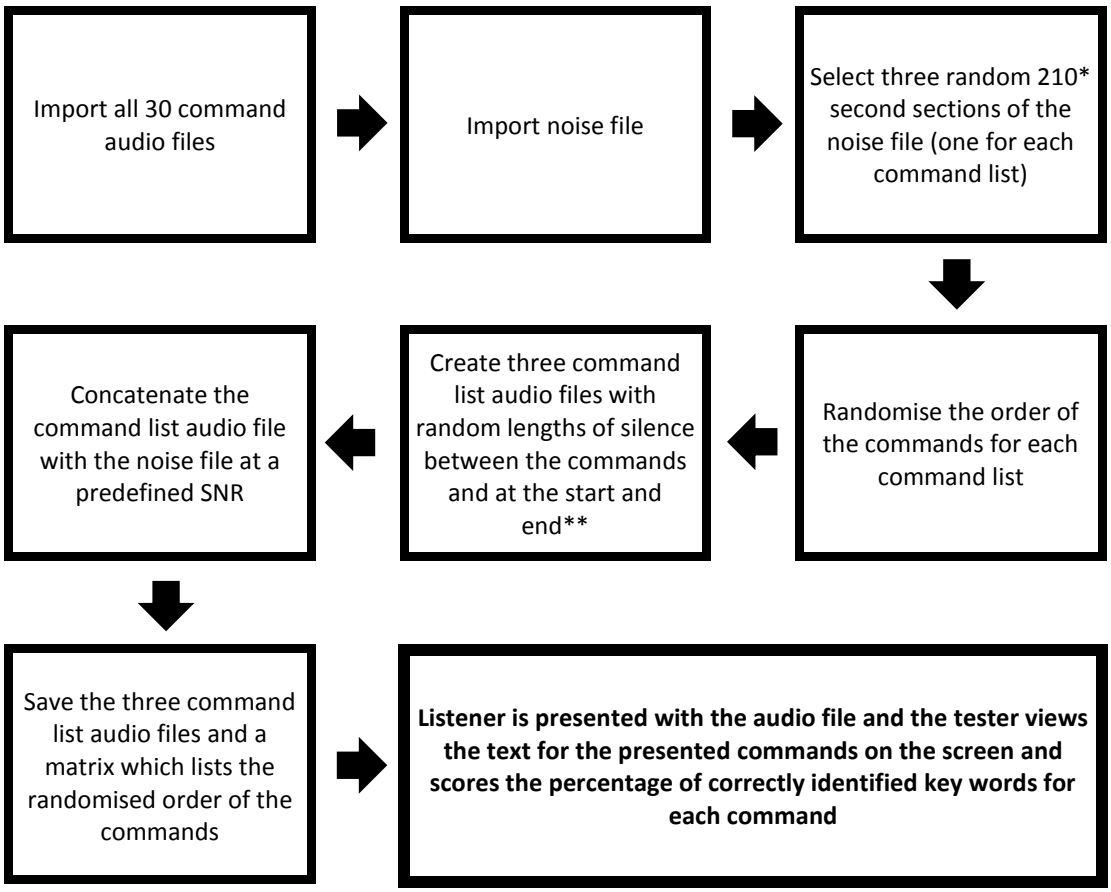
Rather than present all 30 selected commands in one test it was decided that the commands should be split into three lists, allowing for participants to have a short break in between each command list presentation. Each command list contained ten commands and an equal number of key words/phrases.

5.2. *Creating MATLAB code to run the test*

Specific MATLAB (R2013b) code has been written which is designed to generate an audio file concatenating the commands and noise, at a specific SNR, with the commands presented in a randomised order and at random intervals. Figure 6.10 explains the stages of the MATLAB (R2013b) code for developing the command lists ready for presentation to participants.

5.3. *OUTCOME: ready to run VEHCOM SimMCAT*

Stages 5.1 and 5.2 resulted in a ready to run VEHCOM SimMCAT. Figure 6.11 is a schematic diagram representing the audio file presented to a participant listening to one of the command lists.



* 210 seconds allows for enough time for all 10 commands to be presented and adequate time for the participant to repeat it back in between presentations.

** The length of silence between commands is randomised between ten and fifteen seconds. Ten seconds is selected as the shortest time period between commands since the longest presented command is five seconds long, allowing at least double the time for the participant to repeat what they have heard. There is always five seconds of silence before the first command and ten seconds of silence after the last command.

Figure 6.10 *Flow-diagram outlining the stages of the MATLAB code used to prepare the command lists ready for presentation to participants.*

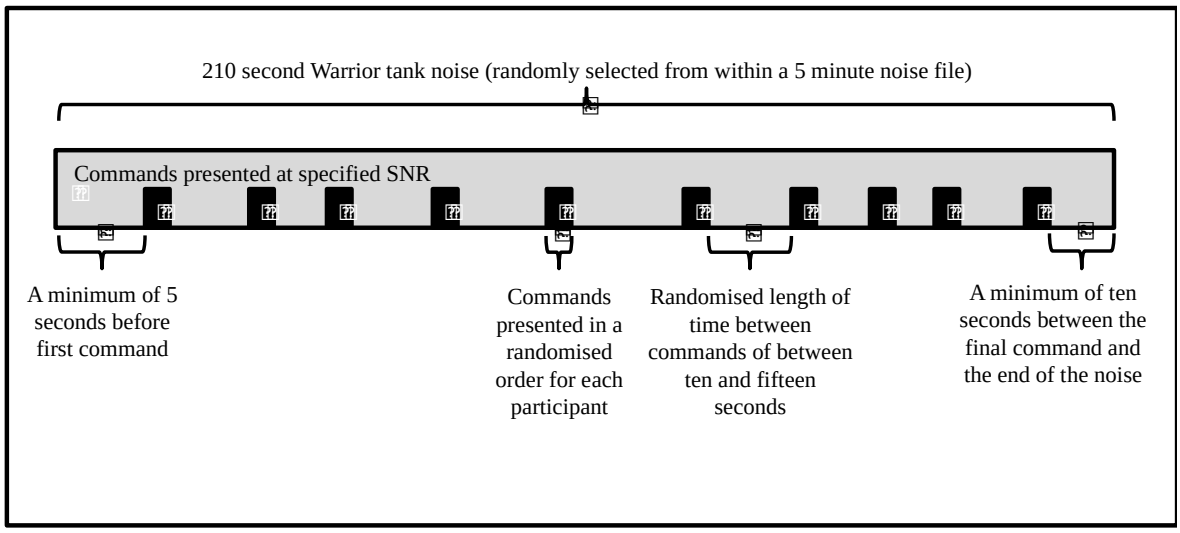


Figure 6.11 *A schematic diagram representing the audio file presented to a participant listening to one of the command lists*

6. Piloting command recognition with normal hearing acuity

A pilot experiment was set up to identify any potential problems within the test methodology, code and scoring method. At this initial stage the piloting was carried out at an advantageous SNR (10 dB SNR), allowing for exploration of how well people are able to listen to and repeat the commands, without the added challenge of adverse listening conditions. The pilot was carried out with a civilian sample of university students.

The stimuli were presented via a SPL Phonitor 2 120 volt audio rail preamplifier connected to an RME Babyface external sound card through Etymōtic ER•2 insert earphones. Calibration was performed in an occluded artificial ear (Brüel and Kjær type 4157, serial number 1901308) using a Brüel and Kjær sound level meter (type 2260 Investigator). The noise was calibrated to be presented at 70 dB A. After only testing a small number of subjects (n=5, civilians, university students) it became evident that participants were struggling to remember and repeat some of the longer commands. Almost all of the participants were reporting that they were able to hear the command clearly but by the end of the longer commands they were struggling to remember the sentence to repeat it. Participants were on average correctly repeating 66% of the key words/phrases (the scores for the five participants were 56%, 61%, 63%, 64% and 79%). Since the VEHCOM SimMCAT is not designed to be a memory test the commands being presented needed to be shortened or simplified. The author systematically went through each command, comparing the scores of each participant. Any command for which the majority of participants identified all the key words/phrases was kept. Commands were cropped if participants consistently only repeated part of it correctly (cropped using Adobe Audition). All of the grid reference commands were shortened to contain either the grid reference or the grid reference and one other piece of

information, such as ‘move now to’ or ‘confirm’. This resulted in no one command containing more than three key pieces of information. The RMS amplitude of the modified commands was equalised to that of the noise and other commands. The command lists were reorganised to ensure that each list contained an equal number of key words.

A second pilot was carried out with the modified command list (n=3, civilians, university students) with different participants to the original pilot; participants reported no difficulty in remembering the commands to repeat them. The mean percentage correct across all three command lists was 89%. No difficulties were identified in the test methodology or the MATLAB code.

6.1. *OUTCOME: Three command lists for the VEHCOM SimMCAT*

The final command lists and key words/phrases to be used in Study 6 are shown in Table 6.2. Each command list contains 23 key words/phrases. Results of the test are given as the percentage of correctly identified key words/phrases averaged across the three command lists. The MATLAB code and test methodology has been piloted and shown to work smoothly.

6.3.4 Finalised VEHCOM SimMCAT

A methodology for assessing performance in a specific scenario environment relating to the SC-MCATs has been developed. The VEHCOM SimMCAT specifically focusses on measuring performance in the environment of listening to commands over a radio whilst moving in a large armoured vehicle. The test involves individuals listening to commands recorded over a PRR in the presence of an armoured vehicle engine noise masker at a fixed SNR. Results are displayed as the percentage of correctly identified key words or phrases averaged across the three command lists.

Table 6.2 *The final command lists and key words/phrases (bold) to be used in the VEHCOM SimMCAT*

COMMAND LIST	COMMAND	Key words
1	I REQUIRE THE **/MEDIC/**AT **/MY LOCATION/**	2
	/CASUALTY/HAS **/GUNSHOT WOUND/**TO **/LEFT ARM/**	3
	/CONFIRM/GRID **/387 489/**GRID 387 489	2
	GRID **/451 667/**GRID 451 667	1
	/MOVE/INTO THE WOOD LINE/AND GO FIRM	3
	/AT THE **FORK/IN** THE **TRACK/GO LEFT/**	3
	/STOP/AT THE **/EDGE OF** THE **BUILDING/**	2
	/ENSURE/YOU **/KEEP IN/**THE **/DEAD GROUND/**	3
	/ENEMY FORCES/IN THE **/HEDGE LINE/**	2
	/GO FIRM/AND **/AWAIT MY ORDERS/**	2
2	WE HAVE **/THREE/CASUALTIES/**	2
	/MOVE NOW/TO GRID /**236 796/**GRID 236 796	2
	/ENEMY FORCE/SEEN DIGGING/IN AN /**IED/**	3
	GRID **/146 787/**GRID 146 787	1
	/GET EYES/ON THE **TARGET AREA/NOW/**	3
	/SLOW DOWN/AND **/KEEP YOUR DISTANCE/**	2
	/PICK UP THE PACE/	1
	/MOVE EAST/AND **/COVER/**THE **/HIGH GROUND/**	3
	/ENEMY FORCES/SEEN/AT THE /**EDGE OF THE BUILDING/**	3
	/MOVE /INTO THE **DITCH/**AND **/GO FIRM/**	3
3	GET THE **/CASUALTY/**ON THE **/STRETCHER/**	2
	/ENEMY FORCES/AT GRID **/917 048/**GRID 917 048	2
	I WANT **/YOUR SECTION/**TO **/FORM/**A **/HASTY DEFENCE/**	3
	GRID **/602 706/**GRID 602 706	1
	/PREPARE TO MOVE/	1
	/LIGHTS OFF/NOW/	2
	WE WILL **/CONDUCT/**A **/SHORT HALT/**AT **/THIS LOCATION/**	3
	/ENEMY FORCES/IN **/TOP WINDOW**, **/AWAIT MY ORDER/**	3
	/HALT, **/ADVANCE/**AND /**BE RECOGNISED/**	3
	/GO FIRM/ON THE **/TRACK/**AND **/OBSERVE TO THE SOUTH/**	3

6.4 Study 6: Initial assessment of the VEHCOM SimMCAT: is it sensitive to simulated hearing impairment and military experience?

6.4.1 Introduction

Study 6 is the final experiment in this thesis. It marks the last step made within the scope of this project towards addressing the objective laid out by the thesis title: developing a measure of AFFD for military personnel. This final stage is somewhat a 'side step' towards achieving this goal. The VEHCOM SimMCAT has been developed as a tool for assessing performance on the SC-MCATs (Section 6.3). Study 6 will focus on exploring performance on the VEHCOM SimMCAT, aiming to provide information about its suitability as a tool for assessing the predictive validity of the CRM and PTA as measures of AFFD. Prior to it being used for this purpose, there are still a number of unknown factors relating to performance on the VEHCOM SimMCAT that need investigating. These can be summarised under the following three topics:

1. The predictive validity of the VEHCOM SimMCAT as a measure of performance on the SC-MCAT remains unknown.
2. The first characteristic of an MCAT is hearing dependency. It was judged by the author that the 17 auditory tasks identified in Study 1 part A require the audition of a sound and cannot be carried out using job experience or other sensory modalities alone; the tasks have therefore been deemed hearing dependent. It therefore follows that performance on a simulation of the MCATs, the VEHCOM SimMCAT, will be influenced by hearing impairment.
3. It is assumed that non-psychoacoustic factors will affect performance on the MCATs (see Section 4.2.2), such as: knowledge of language and vocabulary; age; job experience; confidence in own ability or others' ability; stress levels; competition for attention when carrying out multiple tasks; sensory modality impairments other than hearing loss (e.g. visual impairment affecting lip reading).

It is not possible within the scope of this thesis to explore all of the above issues in relation to the performance of the VEHCOM SimMCAT. Arguably, the predictive validity of the VEHCOM SimMCAT as a measure of real world performance is an abstract concept; quantifying all aspects of 'real world performance' is not possible. Two aspects of the unknown factors listed above have been selected to be explored in the initial stages of developing the VEHCOM SimMCAT. These are explored below, including justification for their selection.

1. Sensitivity to hearing impairment

It is expected that the test will be, to some extent, sensitive to hearing impairment; as hearing acuity decreases so will performance on the test. In order to measure the impact of hearing impairment on performance carrying out the VEHCOM SimMCAT and compare this with performance on the CRM, it has been decided that HLS software will be utilised. There are three key reasons for selecting this approach rather than recruiting hearing impaired listeners:

i. Recruiting normal hearing military personnel (to maintain population validity) is much more achievable within the time frame, based on the author's experience when conducting Study 4.
ii. A sample of hearing impaired military personnel would likely come from a wide range of ranks and roles, with varying ages and levels of experience. Recruiting a sample who are of similar rank and role, are a similar age and have similar amounts of experience will help to minimise the impact of confounding variables in Study 6.This will be simpler to do with a normal hearing population as it would be possible to recruit the entire population from a single regiment who have similar deployment and training experiences.
iii. Utilising HLS allows for more control over the type and severity of hearing impairment being investigated.

2. Influence of job experience

It is expected that several non-psychoacoustic factors are likely to affect performance on the SC-MCATs, (outlined in Section 4.2.2), but it would be difficult to assess whether the VEHCOM SimMCAT is influenced by all of these factors within one experiment. It has been decided that the impact of level experience in operational settings will be explored, specifically focusing on knowledge of language and vocabulary. The VEHCOM SimMCAT is specifically assessing performance on the auditory component of the SC-MCATs; it is thought that performance on these tasks will be influenced by previous experience in operational settings in a way that clinical speech tests, such as the TDT and CRM, are not. This factor has been selected for three reasons:

1. The VEHCOM SimMCAT focusses on assessing the auditory component of performance on the SC-MCAT.
2. In comparison to other non-psychoacoustic factors influencing performance on the SC-MCATs, it is relatively easy to quantify experience in a bimodal manner; 'no experience' e.g. civilian population and 'some experience' e.g. military population.
3. Exploring this particular factor allows for two further issues to be explored which will help inform future AFFD work. Firstly, if military performance is better than civilian, across a

range of hearing acuities, this may suggest that to some extent military personnel are able to compensate for a hearing impairment based on experience, and this should be considered when assessing AFFD. Secondly, if the results from the VEHCOM SimMCAT indicate that military personnel are able to use experience to compensate for reduced hearing acuity, but the performance gap observed on the VEHCOM SimMCAT is not observed on the CRM, this indicates that the CRM is not able to account for the impact of experience on AFFD. If a speech test alone is unable to measure this factor which influences AFFD then additional testing or information gathered from a questionnaire may be required to determine an individual's AFFD status.

In order to explore the two factors mention above, the performance of normal hearing civilian and military personnel will be measured on the CRM and the VEHCOM SimMCAT, when listening through HLS, simulating normal, mild, moderate and severe hearing losses. Assessing whether performance on the VEHCOM SimMCAT decreases as simulated hearing acuity worsens will provide information about whether the VEHCOM SimMCAT is sensitive to hearing impairment. Comparing performance on the CRM between normal hearing individuals listening through the HLS and hearing impaired listeners (using results from Study 4) will provide information about the accuracy of the HLS. In order to explore the influence of job experience on AFFD, differences in performance on the VEHCOM SimMCAT, between the civilian and military personnel, will be compared. It is predicted that if job experience influences AFFD then military personnel will outperform the civilians on the VEHCOM SimMCAT (a military specific task) but not on the CRM.

In Study 4 two scoring methods for the CRM were explored; correctly identifying all three target words (CRM-CSon) or only the colours and number target words (CRM-CSoff). In Study 6, the results from the CRM are being used: 1) to assess the accuracy of the HLS and 2) as a control test to identify whether job experience impacts performance on the VEHCOM SimMCAT (no performance difference is expected on the CRM between civilians and military personnel). It was decided that performance on only one CRM scoring method was required for these goals. This also reduces testing time, which is important due to limited testing time for the military participants, who were only able to commit to short testing sessions. In Study 4 both of the scoring methods were found to display similar levels of measurement precision, not warranting the use of one over another as a more accurate predictor of SRT. It has been decided that the simplest version of the CRM will be run, the CRM-CSoff condition. For simplicity this is referred to as CRM for the remainder of this chapter.

Ultimately, Study 6 will provide additional information about performance on the VEHCOM SimMCAT, which will enable the author to assess its suitability as a tool for assessing performance

on the SC-MCATs and to better understand how to move towards the goal of assessing the predictive validity of the CRM and PTA as measures of AFFD.

6.4.2 Research objective 6 and Study 6 aims

Knowledge gap: The predictive validity of the CRM and PTA as measures of AFFD is currently unknown. In order to assess this, a tool which can measure performance on the SC-MCATs is required. The VEHCOM SimMCAT has been developed as an initial method for achieving this (Study 5). The sensitivity of this tool to hearing impairment and experience is currently unknown.

Research objective 6: to evaluate whether performance on the VEHCOM SimMCAT is affected by hearing impairment and job experience, as would be expected for performance on the SC-MCATs.

Aim 1: To determine if the SC-MCATs are hearing dependent by investigating whether the VEHCOM SimMCAT is sensitive to simulated hearing impairment

Aim 2: To determine whether performance on the VEHCOM SimMCAT is affected by job experience

6.4.3 Hearing loss simulation

Hearing loss simulation (HLS) technology enables the effects of hearing impairment to be studied using a sample of normal hearing listeners presented with stimuli that have been processed in such a way that would, in theory, elicit similar results as would be observed in the hearing impaired population. Section 6.4.1 has explained why it has been decided that this technology will be utilised in Study 6. Section 6.4.3 is split into seven parts: (1) an overview of the chosen HLS; (2) how the HLS works; (3) details about which hearing losses will be simulated; (4) details of how the VEHCOM SimMCAT and CRM stimuli are processed through the HLS; (5) a pilot study to explore performance differences between hearing impaired listeners and normal hearing individuals listening through the HLS on the CRM; (6) a pilot study to select the best SNR to run the VEHCOM SimMCAT experiment which avoids floor and ceiling effects for the different hearing acuities; and finally (7) a summary of Section 6.4.3.

1. Chosen HLS and selected parameters

Key criteria when selecting HLS technology included the principles behind the software being well documented and support being available for manipulating the software for use in Study 6. For the purposes of Study 6 it was not deemed necessary to conduct a thorough review of all available HLSs. In order to explore performance on the VEHCOM SimMCAT and CRM it is necessary to use a

simulator which can, to some extent, indicate the performance levels of hearing impaired individuals on the tests. The Speech Technology Laboratory at the University of Southampton have designed and tested HLS software which is based on the work by Moore and Glasberg (1993) looking at simulating the impact of threshold elevation and loudness recruitment on speech intelligibility in quiet and noise for hearing impaired listeners. The HLS software available at the University of Southampton can be fully supported by the Speech Technology Laboratory team that have developed it, and the principles that it is based on are well documented and accepted (Moore & Glasberg, 1993); it was therefore decided that this software would be used.

The HLS is going to be used to simulate hearing impaired listeners performance on speech intelligibility tasks (performance on the VEHCOM SimMCAT and the CRM). In Chapter 4 it has been discussed that for hearing impaired listeners, factors other than a reduction in audibility contribute towards poorer speech intelligibility performance in comparison to normal hearing listeners. It therefore follows that when simulating hearing impairment, the signal should not only be attenuated but the influence of the additional psychoacoustic abilities that contribute towards speech intelligibility should also be simulated (in the main, frequency selectivity, temporal resolution and loudness recruitment; see Section 4.2.1 for an explanation of these factors and their impact on speech intelligibility). Work is being carried out by the Speech Technology Laboratory to introduce the effect of these factors into the simulator, in addition to simulating the effects of threshold elevation and loudness recruitment using the methods outlined by Moore and Glasberg (1993). However, the simulation of reduced frequency selectivity and temporal resolution is not yet fully developed, in part because the influence of these factors on speech intelligibility is not fully understood. The concepts behind simulating the threshold elevation and loudness recruitment aspects of hearing impairment are much more widely understood, and the work by Moore and Glasberg (1993) is commonly accepted to represent the impact of hearing impairment on speech intelligibility. In Study 6, therefore, the HLS will be used with only the threshold elevation and loudness recruitment elements of hearing impairment being simulated.

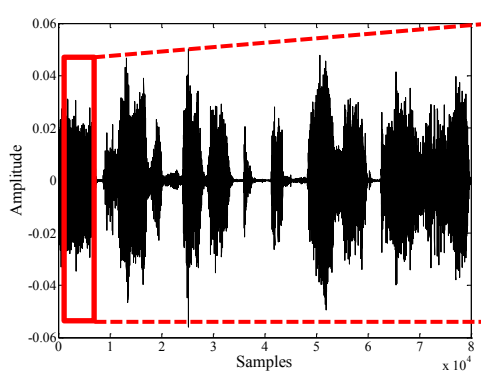
To summarise, the chosen HLS has been developed by the Speech Technology Team at the University of Southampton (main contributor Dr J. Monaghan). The software simulates the loudness recruitment and threshold elevation aspects of hearing impairment and is based on the work by Moore and Glasberg (1993).

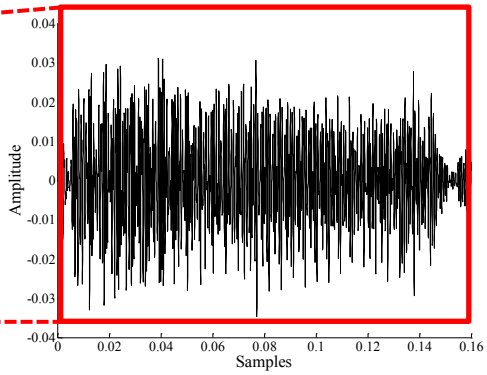
2. An overview of how the HLS works

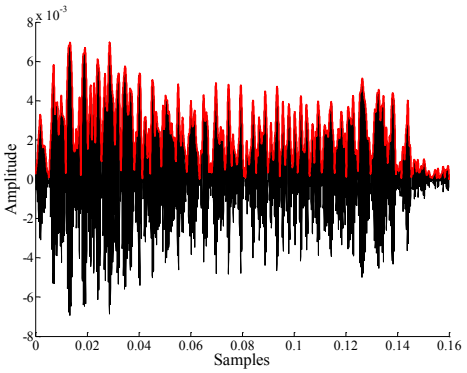
The author acknowledges that the code for running the HLS was designed and written by Dr J. Monaghan, a member of the Speech Technology Laboratory at the University of Southampton. The HLS runs using specifically designed MATLAB (R2013b) code. The stages that the simulator

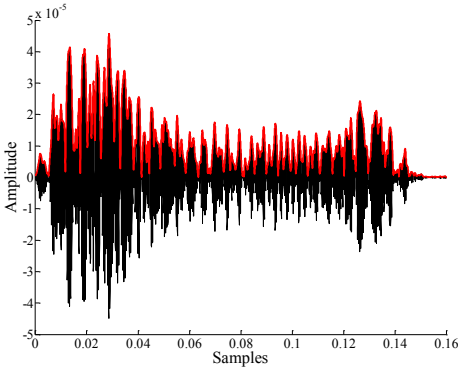
goes through in order to process any audio file are outlined in Table 6.3. The example audio file in Table 6.3 is clean speech. The chosen HLS settings result in two changes being applied to the stimuli; amplitude reduction and loudness recruitment. Figure 6.12 explains how the effects of amplitude reduction and loudness recruitment are simulated and how these changes are applied to the stimuli, relating to Stage 4 in Table 6.3.

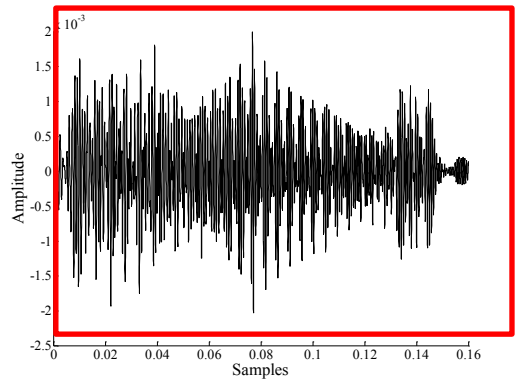
Table 6.3 *Description of how the hearing loss simulator (HLS) works*

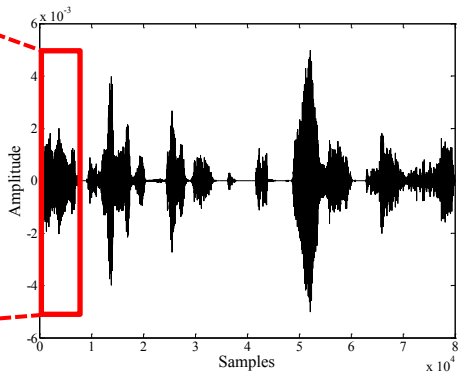
<table>
<tr><td>Stage 1: Audio file imported</td><td>Stage 2: Audio file split into 19 frequency bands</td></tr>
<tr><td> 

Command waveform example</td><td> 

First 160ms of the waveform for one frequency band, centre frequency 2570</td></tr>
<tr><td>Stage 3: A Hilbert transform is applied to the waveform for each frequency channel</td><td>Stage 4: Amplitude reduction and recruitment applied to the envelope (see Figure 6.5 for details). Fine structure multiplied by envelope</td></tr>
<tr><td> 

Hilbert transform of Stage 2 figure (red=envelope, black= fine structure)</td><td> 

Post HLSs Hilbert transform of Stage 3 figure</td></tr>
<tr><td>Stage 5: A new waveform, post HLS is generated for each frequency band</td><td>Stage 6: The waveforms at each frequency band are concatenated, forming the post HLS audio file</td></tr>
<tr><td> 

First 160ms of command waveform post HLS for one frequency band</td><td> 

Whole command waveform post HLS</td></tr>
</table>

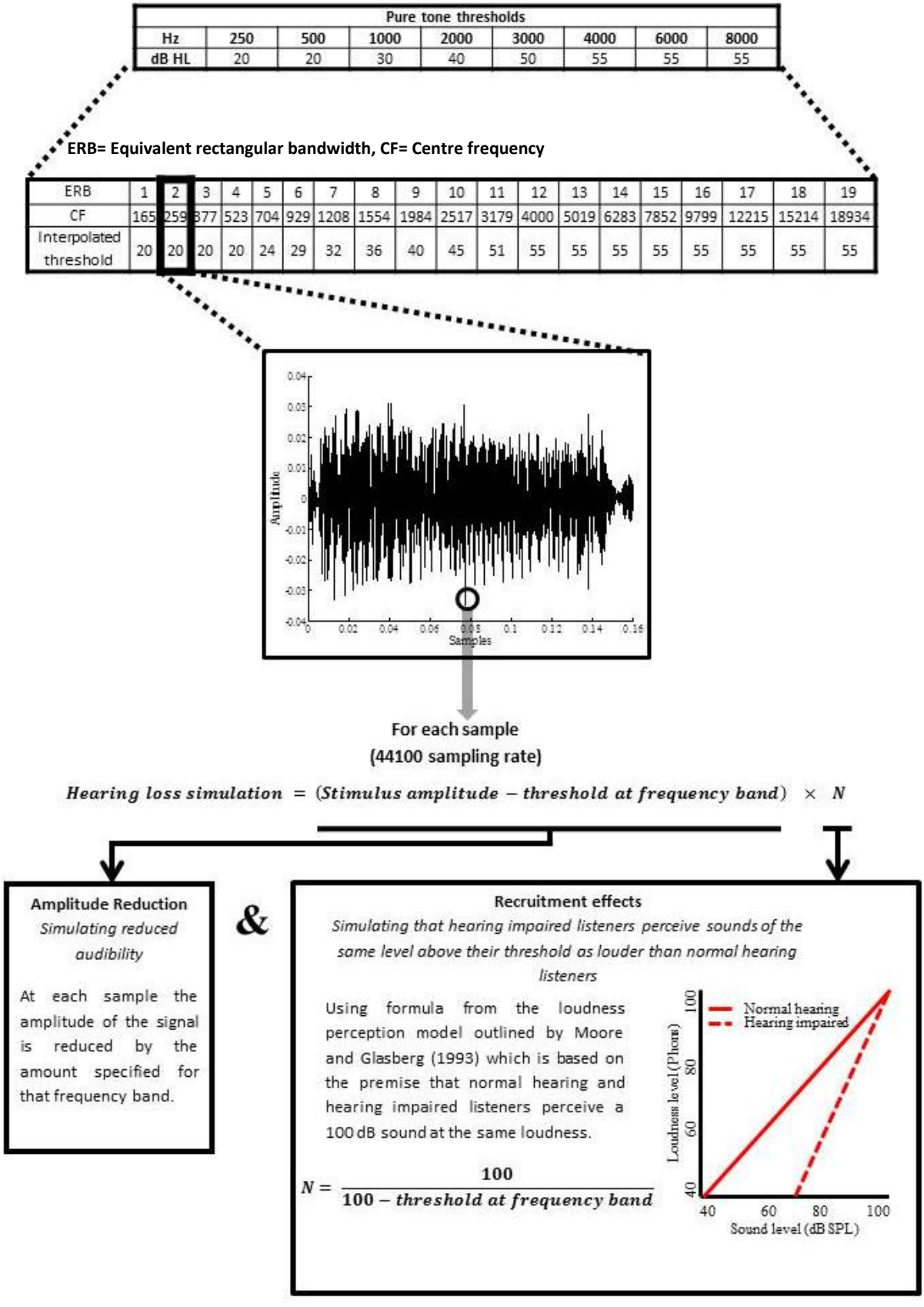


Figure 6.12 *Explanation of how the hearing loss simulation applies the effects of amplitude reduction and loudness recruitment to an audio file*

3. Selecting hearing thresholds to simulate

The final stage of setting up the HLS ready for use in Study 6 is deciding which pure-tone thresholds to simulate. When deciding which hearing acuities to simulate there were two main factors to consider. Firstly, the hearing thresholds should relate to those observed within the military population since this is the population that the results of Study 6 will be applied to, thus increasing the population validity of the study. Secondly, it is of interest to select hearing thresholds which can be related to the H grades (see Chapter 2) currently used to determine AFFD, potentially allowing for some discussion about the suitability of these cut off points.

The Defence Audiology Service (based at the Institute of Naval Medicine) carries out all the audiological care for military personnel who are still in service. In order to gather a picture of the pattern of hearing loss for the military population who are undergoing audiological care of some description, the air conduction thresholds of 400 audiograms were inputted to a single spreadsheet. The audiograms were selected simply by typing in the data of the 400 patients who had most recently attended the clinic (working backwards from the day before data entry). The results for the left and right ears are shown in Figure 6.13. For display purposes the pure-tone thresholds have been averaged across participants for each H grade, for right and left ears separately.

The hearing thresholds observed within the military population were used to influence the hearing thresholds that would be chosen for simulation. The overall shape of the hearing thresholds observed in the military population was maintained for the selected thresholds for simulation, whilst at the same time ensuring that the chosen thresholds fall fairly centrally within each of the H grades. The four audiograms chosen for simulation are shown in Figure 6.14, and Table 6.4 shows where the simulated hearing impairments fall within the H grades. In order to illustrate the similarity between the hearing thresholds chosen for simulation and those observed in the military population they have been plotted on the same audiogram in Figure 6.15, displaying the hearing thresholds of the military personnel taken from the Defence Audiology Service averaged across right and left ears.

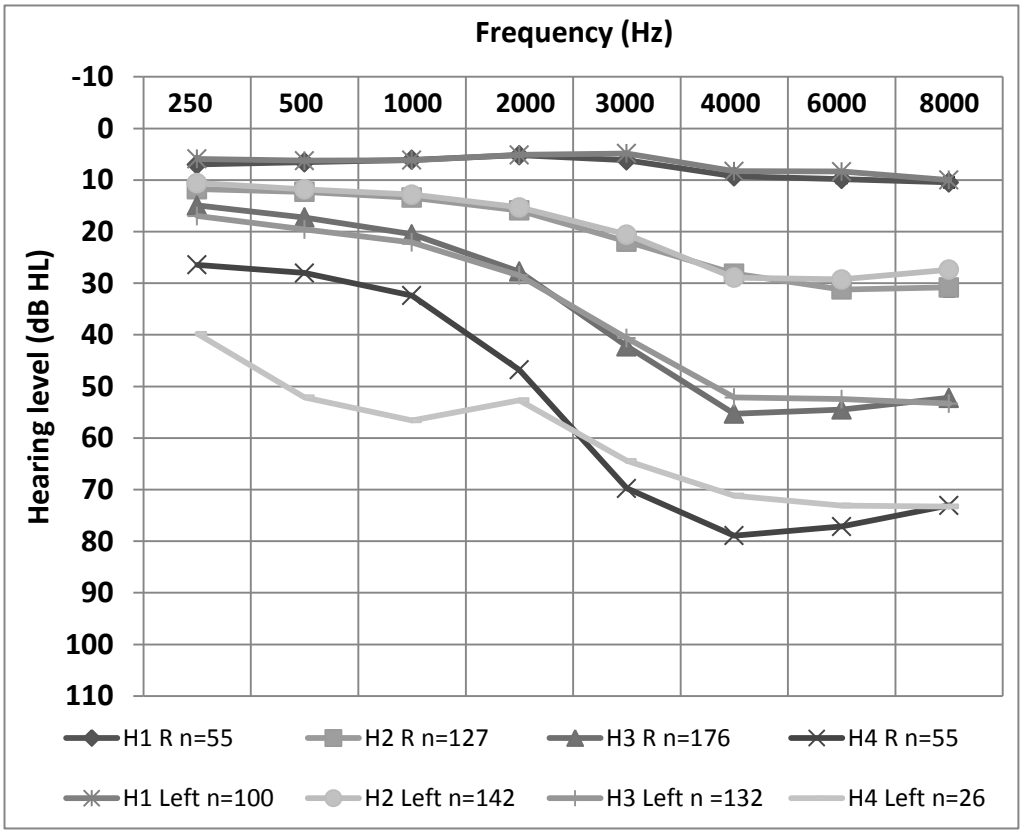


Figure 6.13 *Average air conduction hearing thresholds of 400 patients receiving care at the Defence Audiology Service*

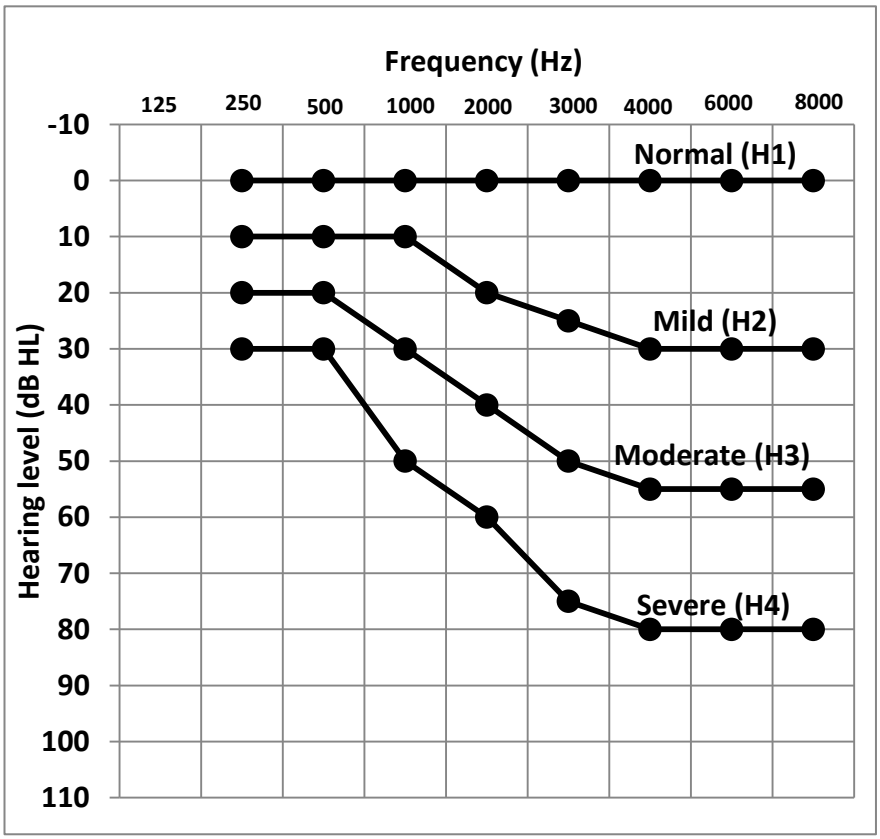


Figure 6.14 *Hearing threshold chosen to be simulated for Study 6*

Table 6.4 *Details of the how the hearing thresholds chosen for simulation fall within the H grades*

	Low frequency (0.5, 1 & 2 kHz) sum (dB HL)	Low frequency H grade (dB HL)	High frequency (3, 4 & 6 kHz) sum (dB HL)	High frequency H grade (dB HL)
Normal (H1)	0	≤ 45	0	≤ 45
Mild (H2)	40	≥ 46 ≤ 84	85	≥ 46 ≤ 123
Moderate (H3)	90	≥ 85 ≤ 150	160	≥ 124 ≤210
Severe (H4)	140	> 150	235	> 211

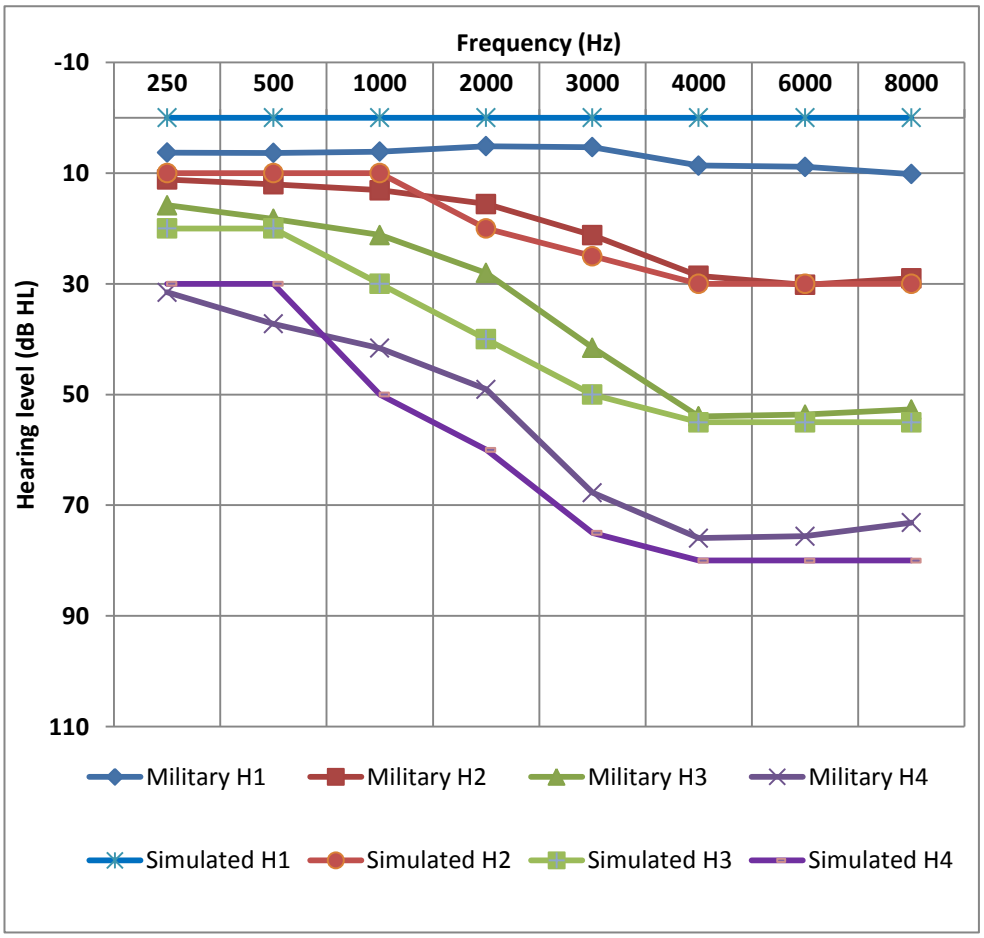


Figure 6.15 *Comparing simulated hearing impairments and those seen in the military population*

4. Running the VEHCOM SimMCAT and CRM adaptive procedure through the HL Sim

For both the VEHCOM SimMCAT and CRM adaptive procedure, the audio files are processed through the HLS at the very last stage before presentation to the listeners, after all other processes have occurred, such as concatenating the sentence and setting the SNR. For the CRM, the adaptive procedure is run in exactly the same way as Study 4, but just before the stimuli are presented to the listener they are processed through the HLS. For the VEHCOM SimMCAT each command list is pre-processed through the HLS prior to presentation to the listener.

The overall presentation level of the VEHCOM SimMCAT and CRM stimuli is determined by two factors, the calibrated level for normal hearing and the simulated hearing acuity levels that defined the amount of amplitude reduction. It is acknowledged that this differs to the method used when presenting CRM stimuli to the hearing impaired listeners in Study 4 since they were able to adjust the overall presentation level so the stimuli were audible to them. The output level of the HLS is dependent on the level the stimuli would be for a normal hearing listener; allowing participants assigned to the impaired hearing acuity groups to alter the presentation level would compromise this. For this reason, participants were not able to alter the presentation level but this was not expected to be problematic since the overall presentation level, at least for the moderate hearing acuity group would still be within the audible range for a normal hearing listener. For the simulated severe hearing acuity group, some of the higher frequency sounds may be beyond the listener's threshold, but this can be considered to be realistic of a severe hearing impairment.

5. Pilot experiment: comparing scores on the CRM for hearing impaired listeners and normal hearing individuals listening through the HLS

In order to check that the HLS is accurately representing the impact of a sensorineural hearing loss on speech recognition, the CRM thresholds obtained from the hearing impaired sample in Study 4 can be directly compared to the thresholds obtained by normal hearing individuals listening to the CRM through the HLS. It is the CRM-CSoff condition which will be investigated in Study 6 (see Section 6.1 for justification). A pilot experiment was conducted with eight participants (students at the University of Southampton) who completed the CRM-CSoff test three times each, once for each simulated hearing acuity group. The CRM adaptive procedure was identical to those used in Study 4, using the call sign off scoring method. Because the CRM was presented through insert headphones and not supra-aural headphones a transform was applied to the sound presented through the inserts to simulate the same frequency response as the supra-aural headphones used in Study 4.

The hearing losses of the participants in Study 4 were assessed to decide which of the hearing acuity groups they fell into (mild, moderate or severe based on the audiograms shown in Figure 6.14) and the average CRM threshold for each of these groups was calculated. The results are shown in Table 6.5 and Figure 6.16.

Table 6.5 *Comparison of mean SRTs on the CRM (call sign off) for hearing impaired individuals and normal hearing individuals listening through the hearing loss simulator*

Hearing acuity group	Simulated		Hearing impaired	
	Sample size	Mean SRT (dB)	Sample size	Mean SRT (dB)
Mild	8	-9.44	12	-8.4
Moderate	8	-7.5	9	-6.3
Severe	8	-2.8	2	-4

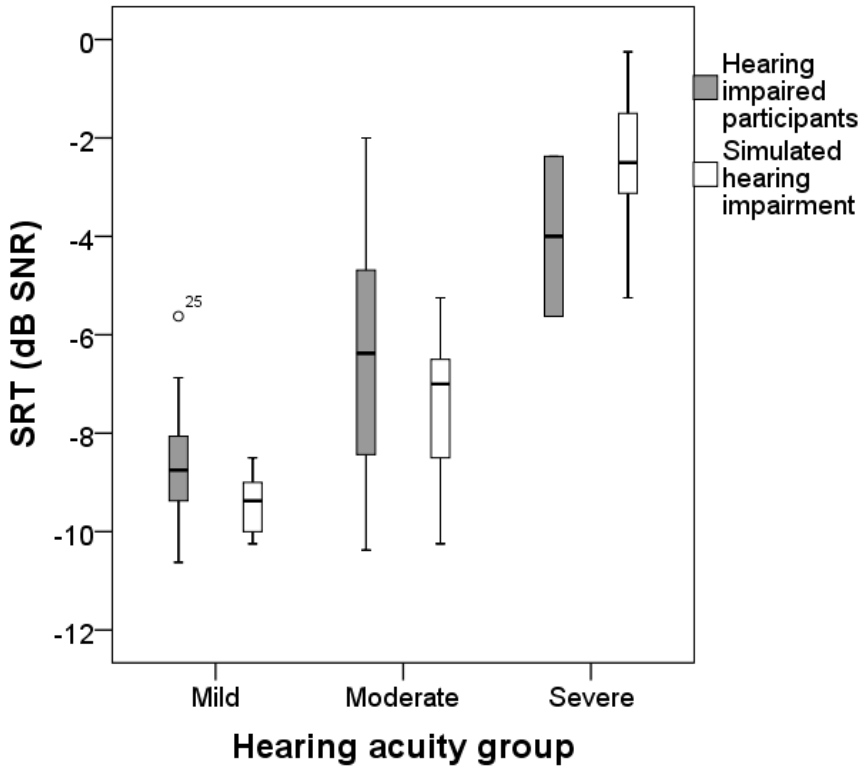


Figure 6.16 *Comparing performance on the CRM for hearing impaired individuals and normal hearing individuals listening through the hearing loss simulator*

The simulated hearing impaired group demonstrated slightly better SRTs than the hearing impaired population for the mild to moderate group. Although this same trend was not observed for the severe group, this result is hard to interpret since there are only two participants in the hearing impaired sample. Considering that only the amplitude reduction and loudness recruitment aspects of hearing impairment are being simulated it is expected that the simulator would underestimate the impact of hearing impairment on speech intelligibility, as observed in the mild and moderate hearing acuity groups. Comparing the severe hearing acuity populations is difficult, since the sample is small, but the difference in scores was not large enough to cause concern this this stage. It was concluded that the results from this pilot provided enough confidence that the simulator is suitable for use in the initial investigation of the VEHCOM SimMCAT.

6.4.4 Method

Military personnel (n=28, mean age=24 years, all male) and civilians (n=28, mean age=24 years, 17 males, 11 females) participated in Study 6. The military participants were tested at Mons Barracks, Aldershot, UK in a quiet classroom. The civilian participants were tested at the Institute of Sound and Vibration Research, Southampton, UK in a sound proof room. All subjects were native speakers of the English language, between 18-40 years of age and had normal hearing. Normal hearing was screened for using PTA (as described by the British Society of Audiology, 2012) and was defined as having hearing thresholds of ≤ 20 dB HL at 0.25, 0.5, 1, 2, 4 and 8 kHz. Ethical approval was obtained for this study from the University of Southampton (ERGO ref: 12943) and MoD Research Ethical Committee (ref: 584).

A sample size calculation has been carried out to ensure that the data is, as a minimum, able to display a significant difference in scores on the VECHOM SimMCAT between the normal and moderate simulated hearing acuity groups. This will allow for exploration as to whether the VEHCOM SimMCAT is sensitive to simulated hearing impairment. Based on the pilot results (Appendix K) a minimum difference in scores on the VECHOM SimMCAT of 15% between the two samples would be of interest, with one standard deviation for each group of 7%. With a significance value of 0.05 and a power of 0.8, a total of 6 participants per hearing acuity group are required (calculated using G*Power software, Faul et al, 2007).

The military personnel were recruited from the Irish Guards who were based at Mons Barracks when the study was being carried out. A Commanding Officer (CO) in 1st Battalion Irish Guards agreed to facilitate the running of the study at Mons Barracks. During the week before the research was carried out at the barracks the CO distributed a flyer and a participant information sheet to the Irish Guards. Personnel who were interested in participating in the study were instructed to report to the training wing on the day the study commenced. The chief investigator (author) then met with all the willing participants and assigned individual dates and times to attend the study. At this stage participants were asked to confirm their participation was voluntary and to complete a consent form. They were reminded that they could withdraw at any time without giving reason and were given the chance to ask any questions.

The civilian sample was recruited from the University of Southampton student population. An email advert was sent to undergraduate and postgraduate students within the Faculty of Engineering and the Environment with details about the study (including the participant information sheet and consent form) and a £20 incentive for participation. Students who replied positively were invited to participate at a time convenient for them. On the day of testing participants were asked to confirm their participation was voluntary and to complete a consent

form. They were reminded that they could withdraw at any time without giving reason and were given the chance to ask any questions.

The experimental procedure for the military and civilian samples was identical. Each participant was assigned to one of four simulated hearing acuity groups, referred to as normal, mild, moderate and severe (see Section 6.4.3 for details about these groups and Figure 6.14 for the corresponding audiograms). Individuals were assigned to a group depending on their subject number. Subjects 1, 5, 9, 13, 17, 21 and 25 were assigned to the normal hearing group and subjects 2, 6, 10, 14, 18, 22 and 26 were assigned to the mild hearing acuity group and so on for the moderate and severe hearing loss groups. Participants were blinded to which group there were assigned to but the tester was not.

All participants, regardless of simulated hearing acuity group, underwent the same experimental procedure. The only difference between the groups was that the stimuli presented to each individual corresponded to the hearing acuity group they were assigned to. The procedure can be split into four stages, shown in Table 6.6. All participants completed the familiarisation stages one and two first, and then odd numbered subjects completed stage three followed by stage four and even numbered subjects completed stage four followed by stage three.

Table 6.6 *Description of Study 6 method stages*

Stage	Order	Explanation of stage
Stage 1	1st	Familiarisation to hearing loss simulator
Stage 2	2nd	Familiarisation to command format and listening over PRR
Stage 3	Odd subjects- 3rd Even subjects- 4th	SRT measurements (CRM-CSoff)
Stage 4	Odd subjects- 4th Even subjects- 3rd	VEHCOM SimMCAT

The experimental stimuli were generated and presented using specifically designed MATLAB (R2013b) code (see Appendix L for details on code authors). The experiment was run using a Mac laptop, running OS X Version 10.9.1. The stimuli were presented via a SPL Phonitor 2 120 volt audio rail preamplifier connected to an RME Babyface external sound card through Etymōtic ER•2 insert earphones. Calibration was performed in an occluded artificial ear (Brüel and Kjær type 4157, serial number 1901308) using a Brüel and Kjær sound level meter (type 2260 Investigator). All calibration was performed using the stimuli that would be presented to the normal hearing acuity listeners. The output level for normal hearing listeners (in dB A) is required by the HLS code in order to adjust the stimuli for the hearing impaired stimuli accordingly.

Stage 1 Familiarisation to hearing loss simulator

It was acknowledged that a single testing session would not be long enough for individuals to become acclimatised to the hearing loss they were being asked to listen to. However, it important that the first time the participants heard a sentence processed through the HLS they were not being scored on that response and they were given an opportunity to familiarise themselves with the HLS. Presenting participants with BKB sentences that have been processed through the HLS for their assigned hearing acuity group provided an opportunity for familiarisation. To ensure all subjects underwent the same testing procedure, subjects assigned to the normal hearing acuity group also listened to these sentences,

Participants were presented with ten sentences from the first BKB sentence list, which were processed through the HLS in accordance with their assigned hearing acuity group. The sentences were calibrated to be presented at an average speech level of 65 dB A for the normal hearing acuity group. They were instructed that the sentences had been processed through a HLS and that the aim of this part of the session was for them to familiarise themselves with what the modified sentences sounds liked. They were given the opportunity to repeat the sentence if they were unsure what they had heard and the researcher informed the participant if they had correctly heard the sentence or not. If they were unable to repeat the sentence the researcher would inform the participant what the sentence was and give them an opportunity to listen to it once more.

The sentences were:

1. The clown had a funny face
2. The car engine is running
3. She cut with her knife
4. The children like strawberries
5. The house had nine rooms
6. They're buying some bread
7. The green tomatoes are small
8. He played with his train
9. The postman shut the gate
10. They're looking at the clock

Stage 2 Familiarisation to command format and listening over PRR

To allow participants an opportunity to familiarise themselves with the format and sound of the commands they listening to eight commands, none of which were the same as those presented in the actual experiment. Participants were asked to repeat back what they heard. They were given the opportunity to repeat the command if they were unsure what they had heard and the researcher informed the participant if they had correctly heard the command or not. If they were

unable to repeat the command the researcher would inform the participant what the command was and give them an opportunity to listen to it once more.

Stage 3 Speech recognition threshold measurements (CRM-CSoff)

Speech recognition thresholds on the CRM were measured using the same method as in Study 4; a two-down one-up adaptive procedure, varying the speech level and taking the mean of the final eight reversals as the SRT. The stationary speech-spectrum noise was the same as that used in previous studies, detailed in Appendix D. The adaptive procedure characteristics are those outlined in Section 5.5.3 (see Figure 5.14 for details). Section 5.5.4, Table 5.9 details the test conditions for the CRM-CSoff condition.

The starting SNR used in Study 6 was 9 dB SNR, the same as that used for the hearing impaired sample in Study 4. For simplicity the same starting SNR was used for the normal hearing sample as the simulated hearing impaired sample in Study 6. The stimuli presented to the normal hearing acuity group had a continuous noise level of 63 dB A (the same as Study 4). The presentation levels of the simulated hearing impaired stimuli were quieter as a result of simulated attenuation.

Before each trial was presented the CRM sentences were processed through the HLS. This took between one and two seconds for each sentence, resulting in a short delay before sentence presentation.

Stage 4 VEHCOM SimMCAT

Participants completed the VEHCOM SimMCAT outlined in Section 6.3. Each participant listened to all three command lists (Table 6.2) in the presence of armoured vehicle engine noise, at -5 dB SNR (see Appendix K for justification for the chosen SNR)**.** The order of command list presentation was randomised for each participant using a Latin square. The command lists were processed through the HLS prior to testing. The stimuli were calibrated to have a continuous noise level of 70 dB A when presented to the normal hearing acuity group. This presentation level was selected as the loudest comfortable listening level for individuals with normal hearing acuity, which is still within the noise exposure limits (in an armoured vehicle the noise level would be much louder and the listener would have control over the PRR volume).

6.4.5 Results

An overview of the results from Study 6 is shown in Table 6.7, displaying the mean scores for each hearing acuity group, across the military and civilian populations for both the CRM and the VEHCOM SimMCAT. The CRM-CSoff results are displayed as a SRT (dB SNR); see Section 5.5.3 for details of how this is calculated. For the VEHCOM SimMCAT the results are displayed as the

number of correctly identified key words/phrases averaged across the three command lists and displayed as a percentage.

Table 6.7 *Overview of averaged results from Study 6 for both the CRM and VEHCOM SimMCAT for both the military and civilian samples*

Hearing acuity	Sample	Mean score across sample	
		CRM-CSoff (dB SNR)	VEHCOM SimMCAT (proportion correct, %)
Normal	Military	-11.2	91
	Civilian	-11.8	85
Mild	Military	-10.1	89
	Civilian	-10.3	81
Moderate	Military	-5.3	86
	Civilian	-5.9	65
Severe	Military	1.5	16
	Civilian	-0.2	11

Prior to any statistical analysis of the data, normality testing has been carried out. The Shapiro-Wilk test is appropriate for small sample sizes (<50 samples) and it tests the null hypothesis that the data is normally distributed (Field, 2005a). Normality testing was carried out separately for the military and civilian samples for each of the hearing acuity groups as well as for the combined hearing acuity groups and the combined military and civilian samples. The results are shown in Table 6.8. The results indicate that parametric tests can be used for any analysis that investigates the hearing acuity groups separately, for both samples, but non-parametric testing is required if the groups are combined.

Table 6.8 *Shapiro-Wilk test of normality for Study 6 data*

Test	Simulated hearing acuity group	Shapiro-Wilk normality test ✓ = p>.05 or ✗ = p<.05	
		Military	Civilian
CRM-CSoff (dB SNR)	Normal	✓	✓
	Mild	✓	✓
	Moderate	✓	✓
	Severe	✓	✓
	Combined hearing acuity groups	✗ (p=.002)	✗ (p=.005)
		Combined military and civilian ✗ (p<.001)	
VEHCOM SimMCAT (%)	Normal	✓	✓
	Mild	✓	✓
	Moderate	✓	✓
	Severe	✓	✓
	Combined hearing acuity groups	✗ (p<.001)	✗ (p<.001)
		Combined military and civilian ✗ (p<.001)	

1. Impact of hearing acuity on CRM and VEHCOM SimMCAT performance

The first part of the Results section will focus on the impact of simulated hearing impairment on the CRM and the VEHCOM SimMCAT, addressing aim one. Figure 6.17 shows the performance levels of both the military and civilian samples on the CRM-CSoff for each hearing acuity group. A visual inspection shows that the two samples display very similar performance levels on the test. In keeping with the findings from Study 4, the CRM is sensitive to hearing impairment; on average, as hearing acuity worsens, so do the SRTs.

One-way ANOVAs were carried out separately for the military and civilian samples. On the CRM-CSoff for both samples, there was a statistically significant difference across hearing acuity groups; military, $F(3,24) = 232.6$, $p < .001$ and civilian, $F(3,24) = 188.7$, $p < .001$. A Tukey post-hoc test, the preferred post-hoc test on a one-way ANOVA (Laerd Statistics, n.d.), was conducted. The Tukey post-hoc test revealed that for both the military and civilian samples there was a significant difference ($p < .002$) between all hearing acuity groups apart from between the normal and mild groups.

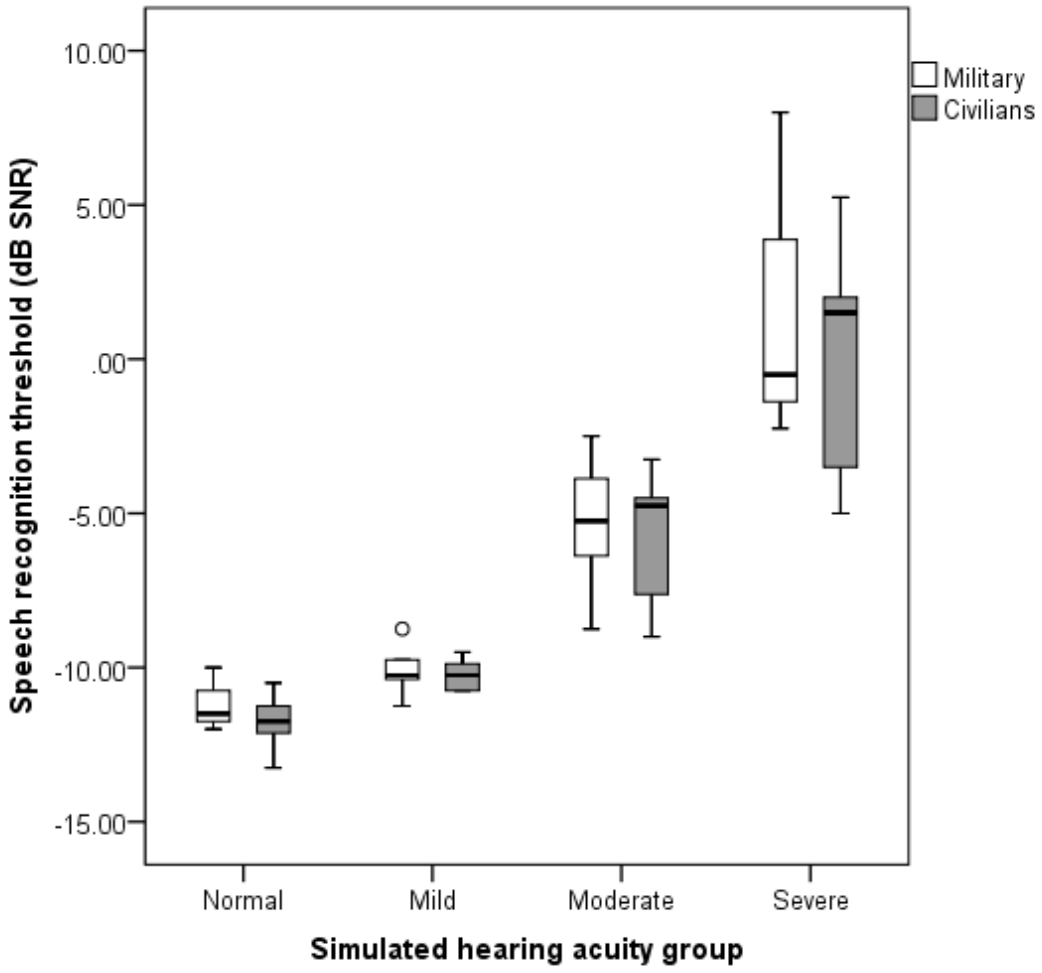


Figure 6.17 *Boxplots showing the performance levels of the military and civilian samples on the CRM-CSoff (outliers exceed 1.5 times the interquartile range)*

In order to understand how accurately the HLS scores relate to the performance of hearing impaired individuals, the CRM-CSoff results from Study 6 (averaged across the military and civilian sample) have been compared with the data collected in Study 4. The data is shown in Table 6.9. Taking into account the standard deviations, performance through the HLS is very similar to that observed in Study 4 for the normal-moderate hearing acuity groups. There is a more noticeable difference of 4.6 dB between studies 4 and 6 for the severe hearing acuity group, but this is difficult to interpret considering the sample size (n=2) of the hearing impaired group. This result may indicate that the simulator is over-estimating the effect of a severe hearing impairment on performance listening to SIN, which will be revisited in conjunction with the results from the VEHCOM SimMCAT.

Table 6.9 *Comparing CRM-CSoff results between hearing impaired listeners and normal hearing individuals listening through the hearing loss simulator*

Hearing acuity group	Study 4			Study 6	
	Sample size	Number of repeats	Mean CRM-CSoff (dB SNR) (1 standard deviation)	Sample size	CRM-CSoff SRT (dB SNR) (1 standard deviation)
Normal	30	4	-10.4 (0.7)	14	-11.5 (0.9)
Mild	12	2	-8.6 (1.4)	14	-10.2 (0.6)
Moderate	10		-6.4 (2.7)	14	-5.6 (2.1)
Severe	2		-4.0 (2.3)	14	0.6 (3.9)

Figure 6.18 shows the performance levels of the both the military and civilian samples on the VEHCOM SimMCAT. On average, as hearing acuity worsens performance on the VEHCOM SimMCAT also worsens, indicating that the test is, to some extent, sensitive to hearing loss. Two one-way ANOVAs were carried out for the military and civilian samples. For both samples, there was a statistically significant difference across hearing acuity groups (military, $F(3,24) = 262.0$, $p < .001$ and civilian, $F(3,24) = 167.8$, $p < .001$). The Tukey post-hoc test revealed that for both the military and civilian samples the severe hearing acuity group scores were significantly worse ($p < .001$) than all other hearing acuity groups. In addition, for both the military and civilian samples the moderate hearing acuity group scores were significantly worse ($p \leq .003$) than the normal hearing acuity group. Differences between all other hearing acuity group pairs were not significant ($p > .500$).

It can be seen that performance on the test deteriorates significantly for the severe hearing acuity group in comparison to the moderate hearing acuity group. At this stage it is not known whether this is a true reflection of how individuals with a severe hearing impairment would perform or whether the HLS over estimates the impact of this level of hearing impairment (as suggested by the results in Table 6.9). This is discussed in Section 6.4.6.

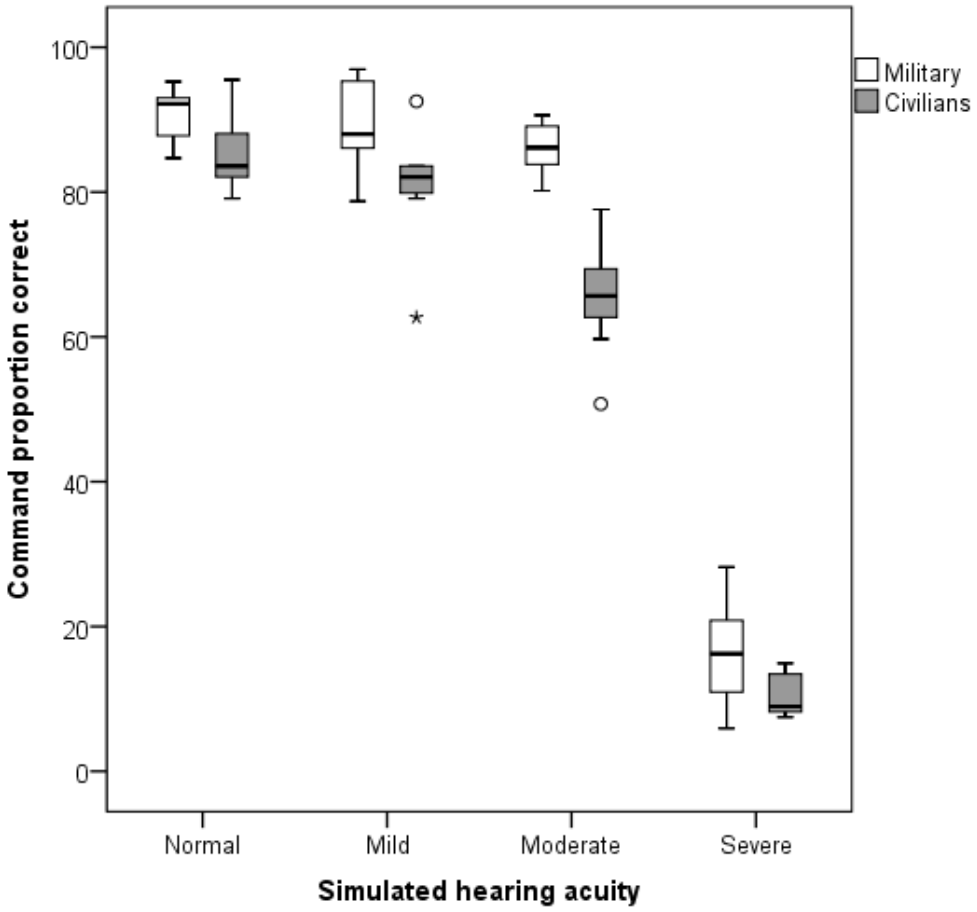


Figure 6.18 *Boxplots showing the performance levels of the military and civilian samples on the VEHCOM SimMCAT (outliers exceed 1.5 times the interquartile range)*

2. Impact of military experience on CRM and VEHCOM SimMCAT performance

The second part of the results section will focus on the impact of experience on performance on the two tests. This is achieved by comparing performance levels between the military (experienced) and civilian (inexperienced) samples and will address aim two.

For the CRM, a visual inspection of the box plots in Figure 6.17 shows that the military and civilian samples display very similar performance levels for all four hearing acuity groups. Four independent-samples t-tests were calculated, to assess for a significant difference between the mean CRM scores of the military and civilian populations for each hearing acuity group. A significant difference ($p<.05$) between the two samples was not found for any of the hearing acuity groups. There is also no trend observed of one sample consistently performing better or worse for each hearing acuity group.

For the VEHCOM SimMCAT, it is evident from looking at Figure 6.18, that the military personnel consistently outperform the civilians; across all four hearing acuity groups the median score for the military population is better than for the civilian population. Four independent-samples t-tests were calculated to assess for a significant difference between the mean VEHCOM SimMCAT

scores of the military and civilian populations for each hearing acuity group. The results are reported in Table 6.10. A significant difference between the two samples was only found for the moderate hearing acuity group ($p < .0001$). It is thought that this performance gap may have also been observed for the normal and mild hearing acuity groups but that performance for the military sample was constrained by ceiling effects; this is further explored in the Discussion section.

Table 6.10 *Independent sample t-tests showing the difference between the VEHCOM SimMCAT scores of the military and civilian populations for each hearing acuity group*

Hearing acuity group	Average VEHCOM SimMCAT proportion correct (%, ± 95% confidence interval)		Independent Samples T-Test (*significant)
	Military	Civilians	
Normal (H1)	90 (±3)	86 (±4)	.08
Mild (H2)	89 (±5)	81 (±6)	.06
Moderate (H3)	86 (±3)	65 (±6)	< .0001*
Severe (H4)	16 (± 6)	11 (± 2)	.11

3. Summary

To summarise, both the CRM-CSoff and the VEHCOM SimMCAT have been shown, to some extent, to be sensitive to a reduction in hearing acuity. No performance differences were observed between the military and civilian sample on the CRM-CSoff, across all four hearing acuity groups. The military sample showed consistently higher scores on the VEHCOM SimMCAT compared to the civilian sample but this difference was only statistically significant for the moderate hearing acuity group.

6.4.6 Discussion and conclusions

Study 6 aimed to evaluate whether performance on the VEHCOM SimMCAT is affected by hearing impairment and job experience. The ultimate goal was to establish whether the VEHCOM SimMCAT is ready for use as a tool for measuring the predictive validity of the CRM and PTA as measures of AFFD and, if not, what next steps are required in order to meet this goal. Within this section three main topics will be addressed:

1. Initial experience running the VEHCOM SimMCAT
2. Is the VEHCOM SimMCAT sensitive to simulated hearing impairment, as would be expected if the MCATs are hearing dependent.
3. What can be concluded about the influence of military experience on the CRM and VEHCOM SimMCAT and, what questions does this raise regarding factors that need to be considered when assessing AFFD?

Initial experience running the VEHCOM SimMCAT

One of the motivations for running Study 6 was to provide information about the initial experience running the VEHCOM SimMCAT and to identify any flaws in the experiment design and test methodology. On a practical note, the author considered the test to be simple and quick to run, requiring minimal specialist equipment and being simple to calibrate. Participants did not demonstrate any difficulty in understanding the test instructions and completed the test as instructed. The main area for discussion with regards to the initial experience of running the VEHCOM SimMCAT relates to the behaviour of the HLS. A decision was made to utilise HLS technology in Study 6, allowing for a larger sample of military personnel to be recruited and for greater control over the configurations of hearing thresholds that were being tested. Although the results from the pilot study detailed in Section 6.4.3 indicated that the HLS was producing SRTs similar to those observed in the hearing impaired population, it is not possible to confirm with any certainty that the performance levels observed in Study 6 accurately represent how a hearing impaired sample would perform. There are two aspects of the results in particular, which are unclear due to the unknown relationship between the HLS and the performance of hearing impaired listeners:

1. The results for the VEHCOM SimMCAT show almost no difference in the scores between the normal and mild simulated hearing acuity groups for both the civilian and military sample. This raises the question as to whether a mild hearing impairment does not impact performance on the VEHCOM SimMCAT and would indicate that this level of impairment may not influence AFFD. However, it is not possible to know whether this finding is a true representation of how a group of hearing impaired listeners with thresholds similar to those simulated would perform. The HLS is only modifying the stimuli by applying amplitude reduction and loudness recruitment to the signal (Section 6.4.3) and therefore does not account for the other psychoacoustic abilities contributing towards reduced speech intelligibility for hearing impaired listeners (reduced frequency selectivity and temporal resolution, see Section 4.2.1). It therefore may be the case that the HLS is overestimating the performance levels of this group and that a larger performance gap would be observed between individuals with normal hearing and those with a mild hearing impairment. A second reason for the similar performance level between these two hearing acuity groups may be due to ceiling effects of the VEHCOM SimMCAT. The test was fixed at a single SNR (-5 dB SNR), which was chosen in order to elicit a range of performance levels amongst the four hearing acuity groups (see Appendix K). However, it may be that a less advantageous SNR would result in a larger performance gap between the normal and mild

simulated hearing acuity groups, and possibly between the mild and moderate groups as well. A consideration for future experimental work would be to consider testing at multiple SNRs (see Section 6.4.7).

2. It is not known whether the steep drop off in performance observed between the moderate and severe simulated hearing acuity groups on the VEHCOM SimMCAT is a true representation of the difference that would be observed between hearing impaired individuals. It may be the case that the HLS over estimates the impact of a severe hearing loss; this deduction would be consistent with the results shown in Table 6.10, where the simulator is shown to over-estimate the impact of a severe hearing loss on performance on the CRM (albeit with a small sample size, n=2, of hearing impaired listeners for comparison). However, it may be that severely hearing impaired individuals would struggle a similar amount on the VEHCOM SimMCAT, suggesting that this group, currently graded H4, should continue to be deemed unfit for duty.

Sensitivity of VEHCOM SimMCAT to simulated hearing impairment

Despite some of the uncertainty relating to the performance of the HLS, the results from Study 6 suggest that the VEHCOM SimMCAT is sensitive to hearing impairment, addressing aim one of this study. This finding is important when considering whether the test is suitable for use as a tool for measuring the predictive validity of the CRM and PTA as measures of AFFD. An assumption was made in Study 1 part B that the MCATs are hearing dependent (Semeraro et al, 2015); it therefore follows that the VEHCOM SimMCAT, designed to simulate the auditory element of the SC-MCATs, is also hearing dependent.

Influence of military experience on performance on the VEHCOM SimMCAT & CRM

The final topic for discussion in this thesis relates to Aim 2 of Study 6; what can be concluded about the influence of military experience on the CRM and VEHCOM SimMCAT, and what questions does this raise regarding factors that need to be considered when assessing AFFD? The results from Study 6 show that there is a performance gap between the military and civilian samples on the VEHCOM SimMCAT (Figure 6.18) and that this performance gap is not observed on the CRM (Figure 6.17). This finding indicates that performance on the VEHCOM SimMCAT is influenced by a variable that is not being measured by the CRM. It is tempting to jump to the conclusion that this variable is 'military experience', but first it is important to consider other confounding variables that may be causing this performance gap between military and civilians. Four factors that the author considers important for discussion are: 1) the influence of hidden hearing loss as a result of noise exposure; 2) differences between the samples in educational

background and intellect; and 3) differences between the samples in cognitive abilities that influence performance on speech intelligibility tasks. These are covered in the following three paragraphs.

It is understood that military personnel are exposed to damaging noise levels (NATO, 2010) and there is evidence to suggest they do not wear hearing protection regularly (Bevis et al, 2014). Animal studies have shown that noise exposure can cause damage to the peripheral and central auditory system that is not evident on the audiogram. For example, Kujawa and Liberman (2009) showed that mice exposed to a 100 dB SPL broadband noise for two hours display significant permanent damage to their inner hair cells and auditory nerve, despite displaying normal behavioural responses to quiet sounds. This is commonly referred to as 'hidden hearing loss' (Plack et al, 2014) and is possibly present in the military sample of Study 6 given their occupational history. There is evidence to suggest that there is a relationship between hidden hearing loss and perceptual deficits in complex listening situations, such as the VEHCOM SimMCAT (Plack et al, 2014). For example, ten audiometrically normal males with a history of noise exposure from sources such as jet engines, small arms fire and helicopter engines, showed a significant deficit in word identification task in noise in comparison to a sample with no history of noise exposure (Alvord, 1983). This evidence would lead to the assumption that the military personnel may demonstrate worse performance levels on both the CRM and the VEHCOM SimMCAT, but this is not consistent with the findings of Study 6. Firstly, the performance gap is only observed on the VEHCOM SimMCAT and secondly (and of most significance) military personnel display better performance on the VEHCOM SimMCAT in comparison to the civilian sample, the opposite of what would be expected for hidden hearing loss.

A second difference between the civilian and military samples is their educational background and reading age. It is assumed that the civilian sample have a higher level of education in comparison to the military sample. The civilian participants have a higher education (HE) degree or are working towards one. It is assumed that the military participants have not attended a HE institute since the average age of military sample was 23.5 and they are not new recruits (i.e. they have not joined post HE). In addition, according to a parliamentary report (Defence Committee, 2003), almost 40% of new recruits to the Army have a reading age of 11 or lower. If these facts are taken as an indicator that the military sample are of lower general scholastic ability or achievement, in comparison to the civilian sample, then it is important to question whether this may have influenced performance on the CRM and VEHCOM SimMCAT. Kidd et al (2007) and Surprenant and Watson (2001) both explored the relationship between general intellectual ability and speech reception threshold. Both studies measured intellect with the SAT-verbal and SAT-mathematical tests and Suprenant and Watson (2001) also looked at individuals' grade point average. Neither

study reports a significant relationship between general scholastic ability and performance listening to syllables, words or sentences in noise. It is therefore concluded that any potential differences in educational background and reading age between the samples are unlikely to impact performance on the CRM or VEHCOM SimMCAT.

Another confounding variable to be explored is whether differences between the two samples in cognitive ability, which may influence performance on speech intelligibility tasks, may have caused the performance gap observed on the VEHCOM SimMCAT. Akeroyd (2008) conducted a review of individual differences in cognitive ability that relate to differences in speech reception. It was concluded that there is "some link" between speech reception and tests of cognitive performance, including: tests of working memory (Gatehouse et al, 2003; Lunner & Sunderwall-Thorén, 2007; Rudner et al, 2008); IQ tests (Humes, 2002; Humes et al 2007); and measures of sensorimotor reaction time (Van Rooij et al, 1989). No tests of cognitive ability were conducted in Study 6. It is possible that, had performance on certain cognitive tests associated with performance on speech intelligibility tasks been assessed, by chance, the military participants would have performed better than the civilian sample. If this was the case, it may be predicted that a performance gap between the samples would have been observed for both the CRM and the VEHCOM SimMCAT, which was not the case in Study 6. Alternatively, due to the more complex nature of the speech stimuli in the VEHCOM SimMCAT, in comparison to the CRM, differences in cognitive ability may have only impacted performance on this test.

Finally, the known difference between the samples is that, contrary to the civilian participants, the military personnel have experience listening to commands and are familiar with the military specific vocabulary used in the VEHCOM SimMCAT. It is understood that top down processing contributes towards a listener's ability to understand speech in adverse listening conditions (Davis and Johnsrude, 2007; Mattys et al, 2012; Zekveld et al, 2006). The top-down process in question here is the listener's ability to use their experience and knowledge of the syntactic and semantic elements of military communication in order to enhance their performance when listening to the VEHCOM SimMCAT commands in an adverse listening situation (i.e. commands over a radio in the presence of engine noise and through a HLS). Parallels can be drawn with evidence that suggests that non-native listeners demonstrate poorer performance when listening to speech in adverse listening scenarios in comparison to native listeners (Mayo et al, 1997; Na´belek and Donahue, 1984; Rogers et al, 2006). In relation to Study 6, the military commands could be considered a language, to an extent, with the military personnel as 'native speakers' and the civilians as 'non-native speakers'. It is considered a reasonable suggestion that the performance difference observed between the military and civilian sample on the VEHCOM SimMCAT is caused by a difference in experience and knowledge of military communication, impacting the top-down

processing contributing to understanding commands in adverse listening situations. This finding suggests that military personnel are, to some extent, able to use their experience and knowledge to compensate for the negative impact of hearing impairment when listening to military specific commands.

The performance gap is not observed between the samples on the CRM, indicating that the influence of experience measured by the VEHCOM SimMCAT is not assessed by the CRM. This has implications when considering the suitability of the CRM as a measure of AFFD; if an individual is able to compensate for a reduction in hearing acuity by using experience and knowledge then this should be factored in when assessing their AFFD. This also applies to other non-psychoacoustic factors not explored in Study 6 which may influence AFFD, such as ability to cope in stressful situations or ability to use other sensory modalities to compensate for a loss of hearing acuity (see Section 4.2.2).

Summary and conclusions

The purpose of Study 6 was to investigate the VEHCOM SimMCAT as a potential tool for measuring the predictive validity of the CRM and PTA as measures of AFFD. It has been established that the VEHCOM SimMCAT is simple to run in terms of the test methodology. Despite the limitations of the HLS, the VEHCOM SimMCAT has been shown, to some extent, to be sensitive to hearing impairment; this is important, since it is designed to simulate the SC-MCATs, which are assumed to be hearing dependent (Bevis and Semeraro et al, 2014 and Semeraro et al, 2015). It is not possible to use the results from Study 5 to draw any definitive conclusions about the exact performance levels of hearing impaired listeners, which could ultimately inform choices about AFFD cut off points; further work with hearing impaired listeners is required for this.

It is expected that performance on the SC-MCATs will be affected by non-psychoacoustic factors (see Chapter 4). Study 6 has explored one of these factors; the results suggest that performance on the VEHCOM SimMCAT is affected by military experience and knowledge of command vocabulary and structure. Military personnel outperform civilians on the VEHCOM SimMCAT but not on the CRM, suggesting that that the CRM is not sensitive to military experience. When assessing AFFD, measurement tools must account for individual ability to compensate for reduced hearing acuity using other abilities, such as experience and knowledge of language. This raises the question as to whether it is possible to use a single tool to predict performance on the SC-MCATs, or whether a battery of tests are required which gather additional information such as previous military experience.

Study 6 has only explored performance on a selection of the SC-MCATs (T1, T2, T4, T5 and T6), in one specific environment, listening to commands over a radio whilst moving in a large armoured vehicle. Although this was a sensible starting point for developing an understanding of developing an AFFD predictor, in order to measure AFFD more generally simulations in a variety of environments relating to the SC-MCATs should be explored. In order to achieve this, simulations relating to the other SC-MCATs could be developed, explored further in Section 6.4.7.

To summarise, Study 6 has provided important information about an initial strategy for assessing performance on the SC-MCATs. It can be to some extent be considered a large scale pilot study for what is still required in order to develop a tool for assessing the predictive validity of the CRM and PTA as measures of AFFD. Suggestions of how to move forward in order to achieve this goal are outlined in Section 6.4.7.

The following concluding statements have been drawn from Study 6:

- The VEHCOM SimMCAT, designed to simulate the auditory element of the SC-MCATs, was found to be sensitive to hearing impairment. It is therefore concluded that the SC-MCATs are hearing dependent (they require the audition of a sound and cannot be carried out using job experience or other sensory modalities alone), a necessary characteristics of a MCAT.
- Military personnel outperform civilians on the VEHCOM SimMCAT, indicating that military personnel are, to some extent, able to use their knowledge and experience of command structure and vocabulary to overcome adverse listening conditions and to compensate for hearing impairment. No performance gap was observed between the military and civilian sample on the CRM, suggesting that the CRM is not sensitive to military experience and cannot predict the impact this has on an individual's AFFD. When assessing AFFD, inclusion of information about an individual's military experience may improve the predictive validity of a tool.

6.4.7 Moving towards assessing the predictive validity of the CRM as a measure of AFFD: next steps

It has been acknowledged in the discussion surrounding Study 6 that there are a number of issues which need addressing in order to move towards using a simulation of the SC-MCATs to assess the predictive validity of the CRM and PTA as measures of AFFD. Three recommendations for further work are outlined below:

1. The VEHCOM SimMCAT measures performance in a specific environment and covers a selection of the SC-MCATs. In order to obtain a more generalisable measure of performance on the SC-MCATs, a SimMCAT which covers a range of environments and all the SC-MCATs should be designed. In addition, increasing the speech corpus would allow for testing at multiple SNRs, avoiding any problems with floor and ceiling effects (although issues relating to equal intelligibility of commands would need to be explored).
2. The VEHCOM SimMCAT was designed to focus on the auditory element of the SC-MCATs. It was found that performance on the test is affected by military experience and it is thought that other non-psychoacoustic factors will also impact AFFD (such as ability to cope in stressful situations, to utilise other sensory modalities, such as sight, to compensate for hearing impairment, or to carry out multiple tasks simultaneously. Designing a SimMCAT which incorporates these factors, measuring AFFD in its entirety, should be considered, although it is acknowledged that the design and analysis of this type of test is complex. Ultimately, this would allow for exploration of the non-psychoacoustic factors that impact AFFD and which of these are predicted by the CRM and PTA, indicating the extent to which these tests can predict an individual's AFFD.
3. Further testing should be conducted with hearing impaired listeners, rather than using HLS, removing any ambiguity about whether performance levels observed are a true reflection of the hearing impaired population.

In order to move this project forward it is proposed that a revised SimMCAT test is developed which measures performance on all of the SC-MCATs, in a variety of environments, still focusing primarily on the auditory element of the tasks. The aim is to develop a test which is fundamentally based on the tried and tested principles of Study 6 (presenting commands in background noise and scoring a response) but is modified to allow for: 1) the test to be run automatically, i.e. the tester is not required to score a response; and 2) the test to be run at more than one SNR, avoiding issues relating to floor and ceiling effects.

A project has been proposed and approved by the Royal Centre for Defence Medicine. In summary this project will develop a tablet based application which will run the CRM, automated PTA and a revised SimMCAT test, as well as gathering information about military experience, age and otological health. The aim of the project is to gather information about the association between performance on the CRM and PTA and performance on the SimMCAT test, indicating the predictive validity of these tools as measures of AFFD. A large sample of normal and hearing impaired military personnel from regiments across the UK will complete the automated test. The revised SimMCAT will assess performance in a number of operational environments and will include an increased corpus of commands which cover all the SC-MCATs. A scoring method which

is based on comprehension (similar to The Listening Comprehension Test, Bowers et al, 2006), rather than repetition, and which is scored automatically, will be developed.

6.5 Chapter 6 Summary

The aim of Chapter 6 was to develop a tool for measuring performance on the SC-MCATs, which can ultimately be used to assess whether the CRM and PTA are capable of predicting AFFD for infantry personnel. In Study 5, the VEHCOM SimMCAT was developed. It measures performance on a selection of the SC-MCATs (T1, T2, T4, T5 and T6), in the specific environment of listening to commands over a radio in a moving vehicle. Prior to using this test to assess the predictive validity of the CRM and PTA, there were a number of unknown factors about performance on the test that needed investigating. Two of these factors were explored in Study 6; it was shown that the VEHCOM SimMCAT is sensitive to simulated hearing impairment and that performance is affected, to some extent, by military experience. Performance on the CRM was not affected by military experience, indicating that the CRM is not sensitive to all factors that contribute towards AFFD. Further work is required to develop a revised SimMCAT test that measures performance in a number of operational environments and covers all the SC-MCATs. This test will then be used to determine whether the CRM or PTA, when combined with additional information such as previous military experience, best predicts performance on the SC-MCAT simulations, and ultimately AFFD.

Chapter 7: Summary, conclusions and future research

7.1 Summary

Occupational hearing standards used within the Armed Forces should accurately predict whether personnel have adequate hearing in order to carry out operational duties safely and effectively. In order to maintain situational awareness, personnel must have sufficient hearing acuity to gather important acoustic cues. Due to the nature of their work and the equipment used, hearing loss is a particular problem for military personnel.

Measures of AFFD should be able to accurately predict performance on hearing critical tasks. Within the military, these tasks are termed mission-critical auditory tasks (MCATs). It therefore follows that any measure of AFFD used within the military should be able to accurately predict performance on MCATs. A gap in knowledge was identified in this area: no work has been completed to identify the hearing dependent tasks carried out within the military and to determine the mission criticality of these tasks.

Following a series of focus groups and a questionnaire with infantry and combat-support personnel, 17 MCATs were identified. Nine of the 17 MCATs were identified as being performed by the majority of ranks and roles, performed either weekly or daily and having either major or critical consequences if performed poorly. Considering the importance of these nine MCATs, the auditory skills required to perform them should be prioritised for representation by a measure of AFFD. It is acknowledged that there are limitations to the list of MCATs identified. Firstly, they are only applicable to the infantry and combat-support subset of the Armed Forces; a similar process should be conducted for other subsets of the military to explore AFFD assessment for other roles. Secondly, the process for identification of the MCATs relied on the recall and opinions of those involved in the focus groups and did not include an observation of personnel carrying out their job. It has subsequently been suggested that stealth awareness should be considered as a potential MCAT. Finally, this process focused on the tasks carried out by personnel and not the environment they are completed in; a better understanding of the environments will be required so that accurate simulations can be developed.

In Chapter 4 it was argued that PTA might be unsuitable as a tool for assessing AFFD given the nature of the MCATs, and this applies to other occupations in which hearing critical tasks are carried out (e.g. driving public transport, firefighting and law enforcement). With a particular

focus on the speech communication MCATs (SC-MCATs) in this thesis, it was proposed that speech intelligibility testing should be explored as an alternative AFFD assessment method.

The CRM SIN test was selected for further investigation, partly due to its high face validity in comparison to command structure, its quickness and simplicity to run and its minimal instructions for the participant or training to run the test, all increasing the likelihood of implementation in the future (should it be found to predict the SC-MCATs adequately). The CRM speech material was re-recorded in British English using NATO call-signs, levels of the stimuli were equalised in terms of intelligibility and it was implemented in an adaptive procedure. The CRM adaptive procedure, in stationary speech-spectrum noise, was shown to have adequate measurement precision to be used to measure individual speech recognition thresholds and there was no difference in the measurement precision between the two CRM scoring methods (CRM-CSon and CRM-CSoff). It is concluded that the CRM adaptive procedure in stationary speech-spectrum noise is a 'ready to use' SIN test, displaying good reliability and concurrent validity for both scoring methods, and is sensitive to military relevant hearing impairment.

In order to evaluate and compare the predictive validity of the CRM and PTA as AFFD assessment methods, a tool for measuring performance on the SC-MCATS is required. The VEHCOM SimMCAT was designed and developed as the first step towards designing a simulation of performance on the SC-MCATs, specifically focusing on the environment of listening to commands over a radio in a moving vehicle.

There were a number of unknown factors about performance on the test that needed investigating prior to using this test to assess the predictive validity of the CRM and PTA. Two of these factors were explored and it was shown that the VEHCOM SimMCAT is sensitive to simulated hearing impairment (confirming that SC-MCATs are hearing dependent) and performance is affected by military experience. It is thought that military personnel may be able, to some extent, able to use their experience and knowledge to compensate for the negative impact of hearing impairment when listening to military specific commands. The CRM-CSoff was not sensitive to military experience and this has implications when considering the suitability of the CRM as a measure of AFFD; if an individual is able to compensate for a reduction in hearing acuity by using experience and knowledge then this should be considered when assessing their AFFD. Further work should explore the predictive validity of the CRM and PTA as AFFD assessment methods and investigate the impact of non-psychoacoustic factors on AFFD and how this should be accounted for when considering assessment tools/test batteries.

It is worth noting here that the work in this thesis has not addressed the very important issue of differentiating between noise-induced hearing loss (NIHL) and high-frequency sloping hearing loss.

In relation to AFFD, these two distinct forms of sensorineural hearing loss have very different audiometric shapes which may influence performance when carrying out MCATs and this is important to consider when designing and evaluating a measure of AFFD. It is not uncommon for the terms 'NIHL' and 'hearing loss' to be used interchangeably when discussing this issue within the military setting, despite their distinctive aetiologies. The guidelines published by Coles et al (2000) are commonly used for diagnosing NIHL and distinguishing between this type of impairment and other types of sensorineural hearing loss. Coles et al (2000) discuss a characteristic downward notch in the 3-6 kHz range that is a typical audiometric configuration for NIHL. This does not concur with averaged shape of the 400 audiograms from military personnel reported in Figure 6.13 (page 176). If it assumed that due to nature of the participants jobs the majority of hearing losses reported in Figure 6.13 are to some extent caused by noise damage this raises a question about the suitability of the widely accepted criteria for defining NIHL. It is suggested that some further work is carried out to investigate the challenges of identifying and diagnosing individual cases of NIHL and how this will influence the assessment of AFFD.

Prior to the work conducted within this thesis, knowledge and understanding of AFFD within the UK Armed Forces was limited. As a result of the work described in this thesis, there is now a clear argument for exploring alternative AFFD assessment methods as well as a recommendation for exploring whether the British English version of the CRM is able to accurately predict performance the SC-MCATs. In summary, a large scale study measuring the association between performance on the CRM, PTA and simulated MCATs, completed by a sample of military personnel with a range of audiometric thresholds, should be prioritised for completion following on from this thesis. The work in this thesis has also prompted a number of additional research projects that the author will be a project investigator on; these are all at various stages of review and are outlined in Section 7.3.

7.2 Conclusions

1) Seventeen MCATs carried out by infantry and combat-support personnel in the British Army have been identified. Nine of these MCATs are performed by the majority of ranks and roles either weekly or daily and have either major or critical consequence if performed poorly (Semeraro et al, 2015). For this subset of the Armed Forces, a measure of AFFD must be able to predict performance on these nine MCATs to ensure they have the necessary auditory skills for safe and effective deployment on operational duties.

2) There is reason to question the suitability of PTA as a tool for assessing AFFD. With a particular focus on the SC-MCATs, there is no clear agreement about the correlation strength between PTA

and measures of speech intelligibility. It is assumed that an equally weak, if not weaker, correlation will be observed between PTA and performance on the SC-MCATs as was reported with speech intelligibility tests. Speech intelligibility tests should be explored as an alternative method for predicting performance on the SC-MCATs.

3) The CRM holds high face validity when compared to command structure and is quick and simple to run; it should be further investigated as a tool for predicting performance on the SC-MCATs. The British English CRM adaptive procedure, in stationary speech-spectrum noise, is a 'ready to use' SIN test. It has been shown to display good reliability and concurrent validity (when compared with the TDT), with two scoring methods (call sign on and call sign off) and to be sensitive to hearing impairment. This test can be used to accurately measure individual SRTs in stationary speech-spectrum noise.

4) The VEHCOM SimMCAT tool has been designed and developed to assess performance on a selection of the MCATs in a specific environment (listening to commands over a radio in a moving vehicle). It is sensitive to simulated hearing impaired and performance on the simulation is, to some extent, affected by military experience. The external validity of the VEHCOM SimMCAT as a tool for assessing performance on SC-MCATs remains unquantified.

5) Military personnel outperform civilians on the VEHCOM SimMCAT, indicating that military personnel are, to some extent, able to use their knowledge and experience of command structure and vocabulary to overcome adverse listening conditions and to compensate for hearing impairment. The influence of this non-psychoacoustic factor (and possibly others that have not been explored here, such as age or working memory) when carrying out MCATs, should be considered when exploring AFFD assessment.

7.3 Future work

1) Identification of HCTs for other occupations and MCATs for other subsets of the Armed Forces.

The methodological framework used by Bevis et al (2014) and Semeraro et al (2015) to identify the MCATs carried out by infantry and combat-support personnel, should be used to investigate the MCATs carried out in other subsets of the Armed Forces (including the Navy and Royal Air Force) and other occupations in which hearing critical tasks are performed (e.g. driving public transport, firefighting, police and manufacturing). In addition to consultation with employees through focus groups and questionnaires, job analysis through observation would provide further information about the tasks carried out within an occupation, without the influence of what the workers deemed to be worth mentioning (a limitation of the methodology in Bevis et al, 2015).

This information should be used for the evaluation of the suitability of AFFD within other occupations where hearing critical tasks are carried out. The author has recently been approached by the UK police force following a recent employment tribunal in which the suitability of PTA was questioned as a tool for measuring AFFD for police officers (The Telegraph, 2015). The author will pursue the potential for future investigation in this field.

2) A systematic review and meta-analysis to answer the question 'to what extent can audiometric thresholds be used to predict speech intelligibility, in quiet and in noise?'

A comprehensive systematic review and meta-analysis should be carried out, including a thorough review of any literature, which reports the correlation between these PTA and speech intelligibility tests. This information is not only important for considering the suitability of PTA as a measure of AFFD but also for supporting further research into a tool for assessing the 'real world listening ability' of hearing impaired listeners in audiology clinics. The results of this meta-analysis would contribute towards addressing priority #9 in the top ten priorities for research on mild-moderate hearing loss; 'How realistic are hearing tests for assessing the everyday hearing abilities of adults with mild to moderate hearing loss?' (Nottingham Hearing Biomedical Research Unit, 2015)

3) Developing a battery of SC-MCATs simulations based on the VEHCOM SimMCAT

The VEHCOM SimMCAT marked an initial attempt towards developing a simulation of the SC-MCATs; it focusses on measuring performance on a subset of the SC-MCATs in a single environment (listening to commands over a radio in a moving vehicle). Using the design and development of this tool as a 'prototype', a battery of simulations should be created which represent performance on all of the MCATs, in a variety of environments, taking into account the recommendations for change in Section 6.4.7. A more detailed investigation of the environments the SC-MCATs are carried out in, either by approaching subject matter experts or running focus groups similar to those conducted in Study 1 part A, would be beneficial prior to creating these simulations. The simulations can be used not only for the investigation of the predictive validity of AFFD assessment tools (covered in point 4) but also for the evaluation of the functional performance of hearing protection devices and radio communication devices.

4) Use SC-MCAT simulations to measure the predictive validity of the CRM and PTA as tools for assessing AFFD

In order to understand whether a test is suitable to be used to assess AFFD, the predictive validity of the tool must be measured. Focusing specifically on speech communication, improved SC-MCAT simulations (point 3) will be used to explore the predictive validity of the CRM (with both scoring methods) and PTA. A project has been proposed and provisionally approved by the Royal

Centre for Defence Medicine. The project involves the development of tablet computer based applications running the CRM, SimMCAT and automated PTA. Tablet computers running these applications will be distributed across multiple regiments throughout the UK and data will be gathered from military personnel to determine the predictive validity of the CRM and PTA, with consideration for the non-psychoacoustic factors (e.g. duration of service, age, rank or role) and how this information could be combined to more accurately predict AFFD.

5) Additional areas to explore: an audit of PTA in the military, hearing protection, tinnitus and the impact of non-psychoacoustic factors on AFFD

Four additional research proposals have been drafted and presented to the Research Council for Defence Medicine (awaiting outcome). The projects are indirectly motivated by the work in this thesis and have been prompted by the authors drive to continue to work in this field. Firstly, concerns have been raised about whether current practice of PTA testing within the military adheres to national standards (British Society of Audiology, 2012) and whether the results are repeatable. It is proposed that a national audit of audiometric testing, leading to national policy recommendations, is carried out. When funding is approved, the author will be involved in the design, logistics and conduction of this study. Secondly, the attitudes of personnel towards hearing protection, a topic raised in the Study 1 part A focus groups (Bevis et al, 2014). It is proposed that a behavioural analysis is conducted in order to understand and improve use of personal hearing protection in infantry personnel. The author is an advisor on this project. Thirdly, the influence of tinnitus on AFFD has not been explored in this thesis. Two projects have been proposed: 1) Does tinnitus impair cognitive function? 2) Development of a digital intervention for the self-management of tinnitus (TINDI) for UK military personnel. The author is an advisor on these projects. Finally, investigating the non-psychoacoustic factors that may influence performance when carrying out the MCATs (such as experience, explored in Study 6, age and ability to cope in stressful situations), and considering how the assessment of these should be incorporated into measures of AFFD, is an important topic to explore. A specific methodology for exploring this has not yet been planned or proposed but the development of improved simulated MCATs will be a first step towards investigating the impact of these factors on AFFD.

Appendices

Appendix A Types of validity

Within this thesis seven types of validity are referred to. To avoid repetition of definitions throughout each chapter a glossary of these terms is provided here.

According to the Oxford English Dictionary (2015) validity can be defined as "the quality of being logically or factually sound". In the context of science and statistics the term 'statistical validity' is often used to encompass all the different types of validity that are relevant to different experiments and refers to the degree to which any conclusions from an experimental study are 'correct' (Shuttleworth, 2009c). Definitions of the seven different types of validity referred to in this thesis are provided in Table A.1. The definitions provided are based on those given by Shuttleworth (2009c).

Type of validity	Definition
Concurrent	This is a measure of how well a test correlates with a previously validated tool measuring the same construct.
Ecological	This is a type of external validity since it refers to the extent to which an effect found in research is generalisable. Ecological validity describes how the testing environment influences experimental results and whether the test environment is representative of the environment the results are being applied to.
External	This refers to the extent to which an effect found in research is generalisable to the environment or population which the results are being extrapolated to represent. Two types of external validity are referred to in this thesis; ecological and population.
Face	This type of validity measures how well a research project 'at face value' relates to the environment or population that the experiment results will be generalised to.
Internal	This refers to the level of confidence that can be placed on a set of test results and involves exploring the level of systematic error within an experiment.
Population	This is a type of external validity since it refers to the extent to which an effect found in research is generalisable. Population validity describes how well the results from the study sample can be extrapolated to the population as a whole.
Predictive	This type of validity refers to how well one test, measuring a certain set of constructs is able to predict performance on another test measuring the same or different set of constructs.

Appendix B Background to psychometric functions (PF)

A psychometric function (PF) is the relationship between a varied stimulus and the likelihood of a particular subjective response. For SIN tests, the stimulus variable is the SNR (plotted on the x axis) and the subjective response is the percentage correct at any given SNR (plotted on the y axis). A sigmoid curve ("S" shaped) is fitted to the data points.

For all of the psychophysical measurement procedures, data is only collected at a small number of the infinite possible stimulus presentation levels. Therefore, the true PF underlying the data is not accessible and needs to be estimated. To do this, it is commonly assumed that the true PF can be described by a specific parametric model and the maximum likelihood estimation is used to estimate the parameters of the model (Żychaluk & Foster, 2009). There are different parametric models used to fit a sigmoid curve to the measured data point. The most commonly used models are: 1) Cumulative normal distribution; 2) Logistic; 3) Weibull; 4) Gumbel; and 5) Hyperbolic secant (Kingdom & Prins, 2010). Once one of these models has been used to fit a PF, predictions of performance at other stimulus presentations can be made and variables, such as the speech recognition threshold 50% (SRT 50), can be obtained. All of these models attempt to construct a curve that has the best fit to the data points using the least squares method is used (Weisstein, 2014).

There is no set method for selecting which model should be applied to a given data set; most commonly the model that displays the least deviation from the measured data points is chosen. This is referred to as the 'goodness-of-fit' and is a measure of the amount a PF deviates from the data. Using the Palamedes toolbox (Prins & Kingdom, 2009), this measure is given by a pDev score, which is always a number between 0 and 1. According to Kingdom and Prins (2010), a pDev score of less than 0.05 indicates that either the quality of the data is poor or the fit of the curve is unacceptably poor and an alternative model should be used.

In relation to SIN testing the model most commonly fitted to this type of data is a logistic function (HearCom, 2006; Leek, 2001; Pedersen & Juhl, 2014; Johansson & Arlinger, 2002; Versfeld & Dreschler, 2002; Rhebergen & Versfeld, 2005). A logistic function was fitted to the data gathered in Study 2 and Study 3 and this model was shown to fit the data well, displaying pDev scores of >0.6. Logistics functions have been used throughout this thesis as the chosen model for fitting PFs.

Once a PF has been plotted, regardless of the chosen model, different pieces of information about speech intelligibility can be obtained. There are four key features of any PF, shown in Figure A.1: 1)

alpha (α) or *location*; 2) beta (β) or *slope*; 3) lower asymptote (γ) or *guess rate*; 4) upper asymptote (λ) or *lapse rate*.

B.1 Location

The alpha (α) or *location* is the overall position of the curve along the x axis when the proportion of correct responses reaches a set criterion. In the context of speech-in-noise (SIN) testing, the location is commonly described as the SNR at which an individual scores 50% correct, or the 50% speech recognition threshold (SRT 50).

B.2 Slope

The beta (β) or *slope* defines the gradient of the curve. No single value can be given to describe the slope of an entire PF, since the slope is changeable across the plot. Typically the value of β is the steepest point of a PF, known as the inflection point. This is the point at which a slope changes direction. A straight line can be plotted at the point of inflection, known as the tangent line. The slope of the tangent line is equivalent to the exact slope value at the point of inflection and is a theoretical concept. To calculate this value, the derivative of a secant line is used (see Figure A.2). A secant line is plotted between the point of inflection and a second 'close' point on the PF (s, Figure A.2). The slope of the secant line gives and approximation of the true value of the slope at the inflection point; the closer s is to the inflection point, or the closer h (see Figure A.2) is to zero, the closer the slope of the secant line is to the tangent line and the better the approximation of the slope. To find the slope of the tangent line the limit of the slope of the secant line (as h gets closer to zero) is calculated by finding the derivative of the slope of the secant line, the equation for which is outlined in Figure A.2 (Khan Academy, 2014). This value describes the slope at the point of inflection. Shallow slopes have lower slope values than steeper slopes; for a shallow slope the difference between $(f)a + h$ and $(f)a$ is smaller, resulting in a lower y: x ratio and therefore smaller slope value.

When a PF is fitted to the data the 'best fit' is used, which can only provide estimates of the threshold and slope parameters. The symbols $\hat{\alpha}$ and $\hat{\beta}$ are used to express that only estimates of these values are displayed. The method of bootstrap analysis allows for estimation of the error for the $\hat{\alpha}$ and $\hat{\beta}$ values and calculates the 95% confidence intervals. A random set of hypothetical data, based on the recorded experimental data, is generated. For each set of hypothetical data, a logistic function is fitted and estimates of the threshold and slope are generated. Using the simulated sets, the estimated 95% confidence intervals of $\hat{\alpha}$ and $\hat{\beta}$ are calculated

B.3 Guess rate

The lower asymptote or *guess rate* (γ) is not usually a 'free' parameter of a PF because determined by the number of options available in the forced choice method. For example a two-alternative-forced-choice (2AFC) method would result in a lower asymptote of 50% since there is a 50% chance of a correct response for every presentation. For the CRM, the lower asymptote for the individual target words are, guess rate for the target words individually are 11.1% for the colours and numbers (nine options for each, 9/100), and 5.6% for the callsigns (18 options for each, 18/100)/. There are 1458 possible CRM sentences so the lower asymptote for guessing the entire sentence correctly is 0.07%.

B.4 Lapse rate

The upper asymptote or *lapse rate* (λ) is determined by the deviation from 100% correct at a presentation level at which consistently correct responses would be expected, giving a description of 'non-perfect' performance. Unlike the guess rate, the lapse rate varies across repeats both between and within subjects. The lapse rate is usually a free parameter but is often constrained by the test method to avoid a test continuing indefinitely as a result of a series of incorrect responses at advantageous presentation levels. For example, for the CRM adaptive procedure (Section 5.5, Study 4) includes a maximum presentation level of 10 dB SNR and termination criteria if a participant continuously responds incorrectly at high SNRs.

B.5 Deriving scores from a PF

There are two main scores that are typically derived from a PF: 1) a percentage correct score at a specific presentation level (a fixed SNR); and 2) the Speech Recognition Threshold at the 50% correct level (SRT 50). Percentage correct scores give an indication of the subjective response at any given stimuli level. SRT 50 scores can be obtained when an intelligibility function has been plotted across the range of frequencies from the guess rate to near 100% correct and the 50% correct SNR is extrapolated.

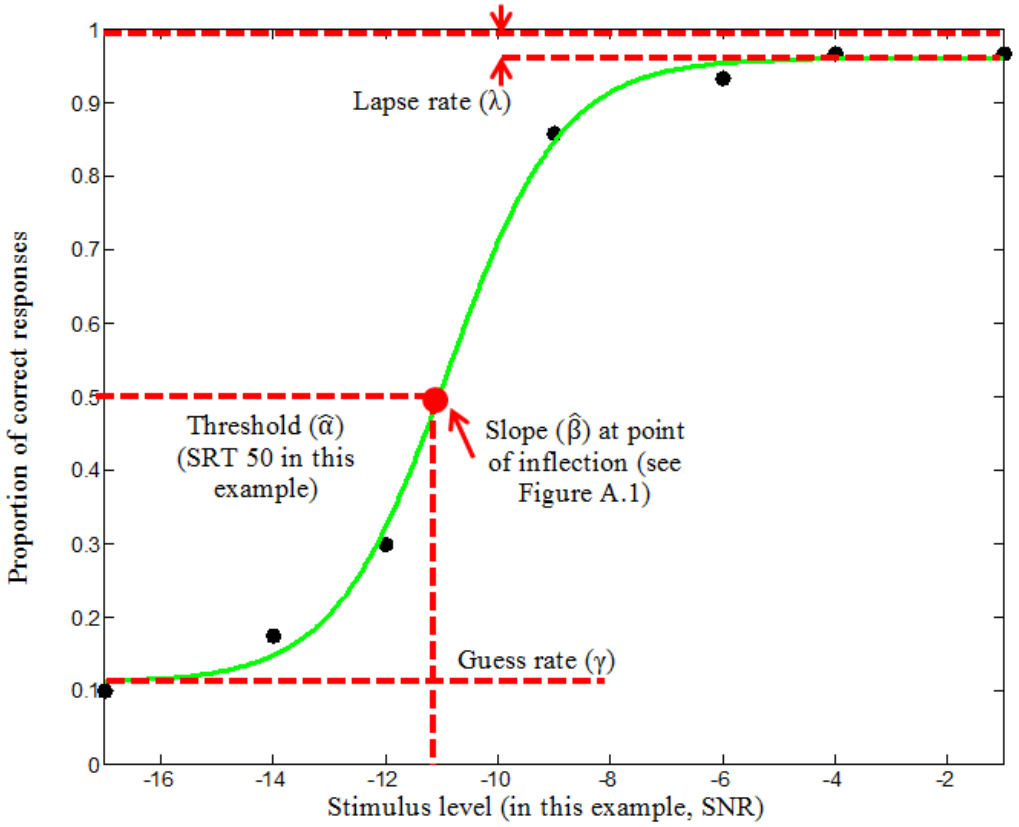


Figure A.1 *Diagram showing points of interest on a psychometric function*

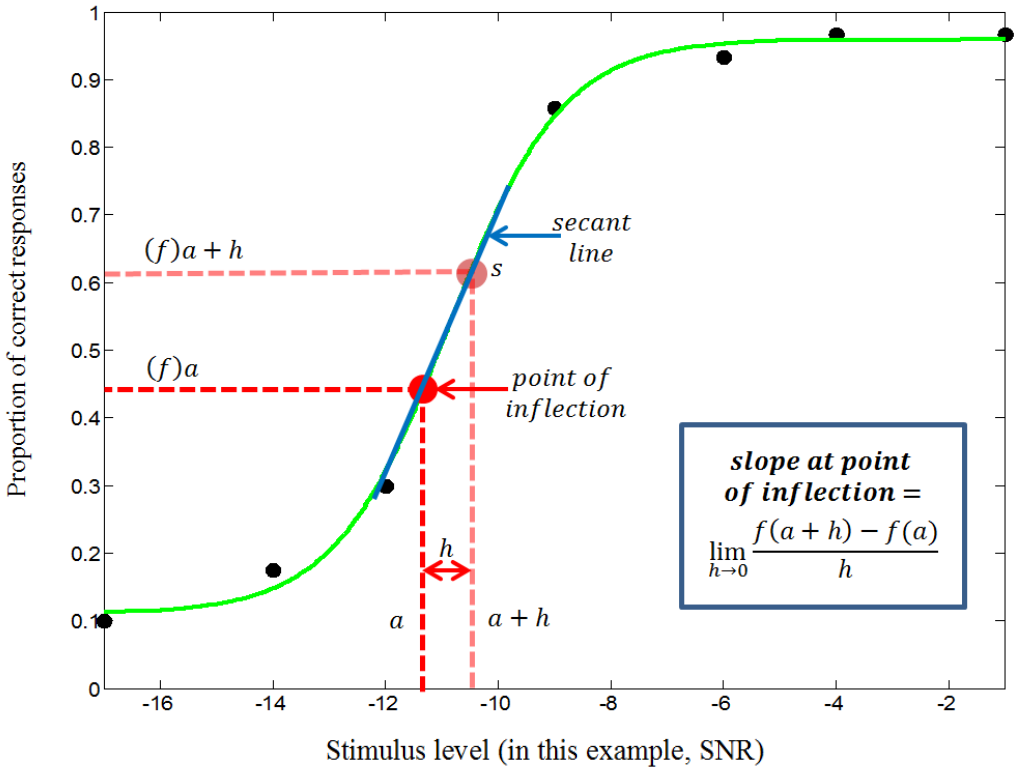


Figure A.2 *Diagram to show slope calculation*

Appendix C Survey to obtain the opinions of military personnel on speech tests

Hannah Semeraro (PhD Student, University of Southampton)
Email: hds1g08@soton.ac.uk
Institute of Sound and Vibration Research
University of Southampton
SO17 1BJ
Tel. 07595675454

STUDY TITLE- DEVELOPING A SPEECH-IN-NOISE TEST FOR MILITARY PERSONNEL

This survey aims to obtain opinions about the relevance of different speech tests to military communication. The speech tests listed below are commonly used in audiology clinics to assess ability to understand speech in the presence of background noise. We are aiming to develop a test which assesses the ability of military personnel to communicate effectively during operations. In the table below are examples of the speech material which are used in different speech tests. Your task is to rank each test in terms of its relevance to military communications, using the scale provided and to use the final column to give a brief justification for your answer. Please read through all of the speech test examples before completing the questions.

Fill in the form electronically by answering the questions in the space provided. Return the form electronically as email attachment to Hannah Semeraro- hds1g08@soton.ac.uk.

Speech test name and examples	**Relevance Rank:** **1= Very relevant to military communications** **2= Has some relevance to military communications** **3= No relevance to military communications**	**Justification for answer**
Bamford-Kowel-Bamford (BKB)		
The participant is tested on how many of the underlined words they can hear and correctly repeat. The clown had a funny face. The car engine's running. She cut with her knife.		
Coordinate Response Measure (CRM)		
The participant is tested on how many of the underlined words they can hear and correctly repeat.		

Ready Tiger go to blue five now Ready Arrow go to red eight now Ready Charlie go to white two now		
Hearing in Noise Test (HINT)		
The participant is required to correctly identify the whole sentence. The boy fell from the window. The wife helped her husband. Big dogs can be dangerous.		
Modified Rhyme Test (MRT)		
The sentence "You will mark (key word) please" is said. The participant is faced with a set of six rhyming key words. They are required to correctly identify the key word spoken. *Example of rhyming key word sets include.* Went, sent, bent, dent, tent, rent. Hold, cold, told, fold, sold, gold. Pat, pad, pan, path, pack, pass.		
Quick Speech-in-Noise Test (QuickSIN)		
The participant is tested on how many of the underlined words they can hear and correctly repeat. The lake sparkled in the red hot sun. Tend the sheep while the dog wanders. Take two shares as a fair profit.		
Triple Digit Test (TDT)		
The participant is required to correctly identify the underlined words. The numbers two, five, three. The numbers eight, one, four. The numbers two, nine, five.		
Words in Noise Test (WIN)		
The sentence "Say the word (key word)" is said. The participant is required to repeat the key word that is said in the sentence. The subject is not *given any words to choose from.* *Example key words are:* Food, pain, late, dodge, cool, ditch, kick, luck, gun, such.		

Appendix D MATLAB code for generating stationary speech-spectrum noise

The following MATLAB (R2013b) code was used to generate stationary speech-spectrum noise that has the same frequency shaping as the CRM sentences. This code was produced by Dr Rachel van Besouw, University of Southampton.

```
%read in the wav file containing the speech
file = uigetfile('*.wav', 'Select the wave file');
[stimuli,fs] = wavread(file);
%normalise using RMS amplitude
stimuli = stimuli/sqrt((sum(stimuli.^2))/length(stimuli));
%generate filter coefficients using LTA spectrum
[a,b]= lpc(stimuli,6);
%generate white noise the same length as the noise =
randn(length(stimuli),1);
%generate speech-shaped noise
speechshapednoise = filter(b,a,noise);
%write output wav file
wavwrite(speechshapednoise, fs, 'speechshapednoise.wav');
```

A wav file was created which contained all of the CRM target words with all silence between the sentences removed (less than 0.1 ms of silence between target words). The frequency spectrum of this file was then analysed and used to generate white noise with the same frequency spectrum. Figure A.3 shows the similarity between the frequency shaping of the noise file (bottom, green) and of the CRM sentence components (top, red). For visual purposes the amplitude of the noise (bottom) has been made approximately 3 dB lower than the CRM sentences (top), so the lines do not overlap. The noise file was 28 seconds long, the same length as the CRM sentence component file.

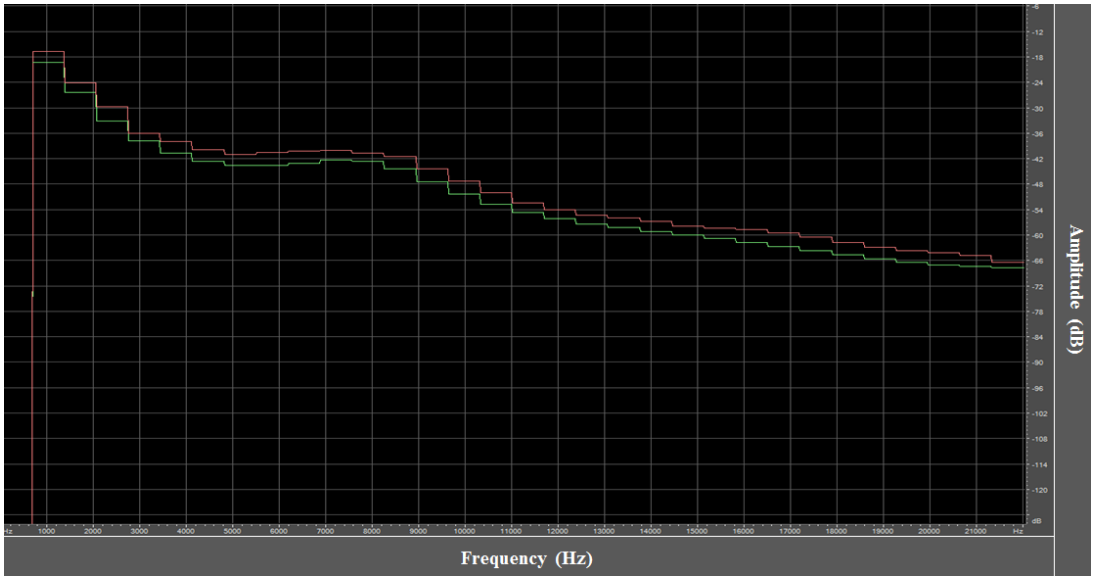


Figure A.3 *Frequency shaping of CRM sentences (top, red) and stationary speech-spectrum noise (bottom, green)*

Appendix E MATLAB code equalising the RMS of the CRM speech stimuli

All the recordings (36 in total: 18 'Ready *call sign*', 9 'go to *colour*' and 9 'number *now*') were processed to have the same RMS amplitude using the following MATLAB (R2013b) code.

```
BaselineRMS = 0.0862;                    % RMS value all sound files will be set
                                         to (chosen by using the original RMS of
                                         the 'Ready Alpha' recording

file_rms = rms(file);                    % finds out the RMS of the selected file

file_sf = BaselineRMS/file_rms;          % sf = scaling factor (diff between
                                         baseline and selected file RMS)

file_rmseq = file*file_sf;               %scaling selected sound file so the RMS
                                         is equal to the baseline RMS

wavwrite (file_rmseq, fs, 'filename');   %writes new sound file as selected
                                         'filename'
```

This was then cross checked by checking the RMS levels of all the sound files in MATLAB and Adobe Audition

.

Appendix F MATLAB code CRM method of constant stimuli

```
      x_signal = [silence_1 ; CS{CS_List(n)} ; silence_2 ; C{C_List(n)} ;
N{N_List(n)} ; silence_1 ]; %concatenates the sentence, with silence
before, between CS and C and at the end

    N_signal = size(x_signal,1); % finds out the size of the signal, to
inform the size of the noise

    index = randi([1 N_noise-N_signal+1],1); % selects a random starting
point for the noise file which is atleast N_signal smaller than the size
of the noise file

    x_noise = noise(index:index+N_signal-1); % creates x_noise which is
the same length as N_signal

    T_RF = 0.15; % Rise and Fall Time in s (same as used in Smits et al
2004 paper)

    T_RF_S = T_RF*fs; % Rise and Fall Time in Samples

   T_Const = size(x_noise,1)-2*T_RF_S; % defines the time that the noise
stays the same (given as ones in ramping vector)

    Ramps = [(0:1/(T_RF_S-1):1).' ; ones(T_Const,1) ; (1:-1/(T_RF_S-
1):0).']; % Ramping vector. (note 350ms of noise at max presentation
level prior to speech starting to prevent forward masking (minimum 200ms
needed)

    x_noise = x_noise.*Ramps;

    RMS_noise = rms(x_noise); % finds out the RMS of the noise (which is
the same as the signal RMS)

    Gain = RMS_noise/RMS_signal*10^(snr/20); % finds out how much gain
needs to be added/removed from the signal to satisfy the selected SNR

    % write final output signal to be played back

    output = Gain*x_signal + x_noise;

    % play the output signal

    sound(output*0.1, fs);
```

Appendix G Graphs showing direct comparison of psychometric functions before and after intelligibility equalisation

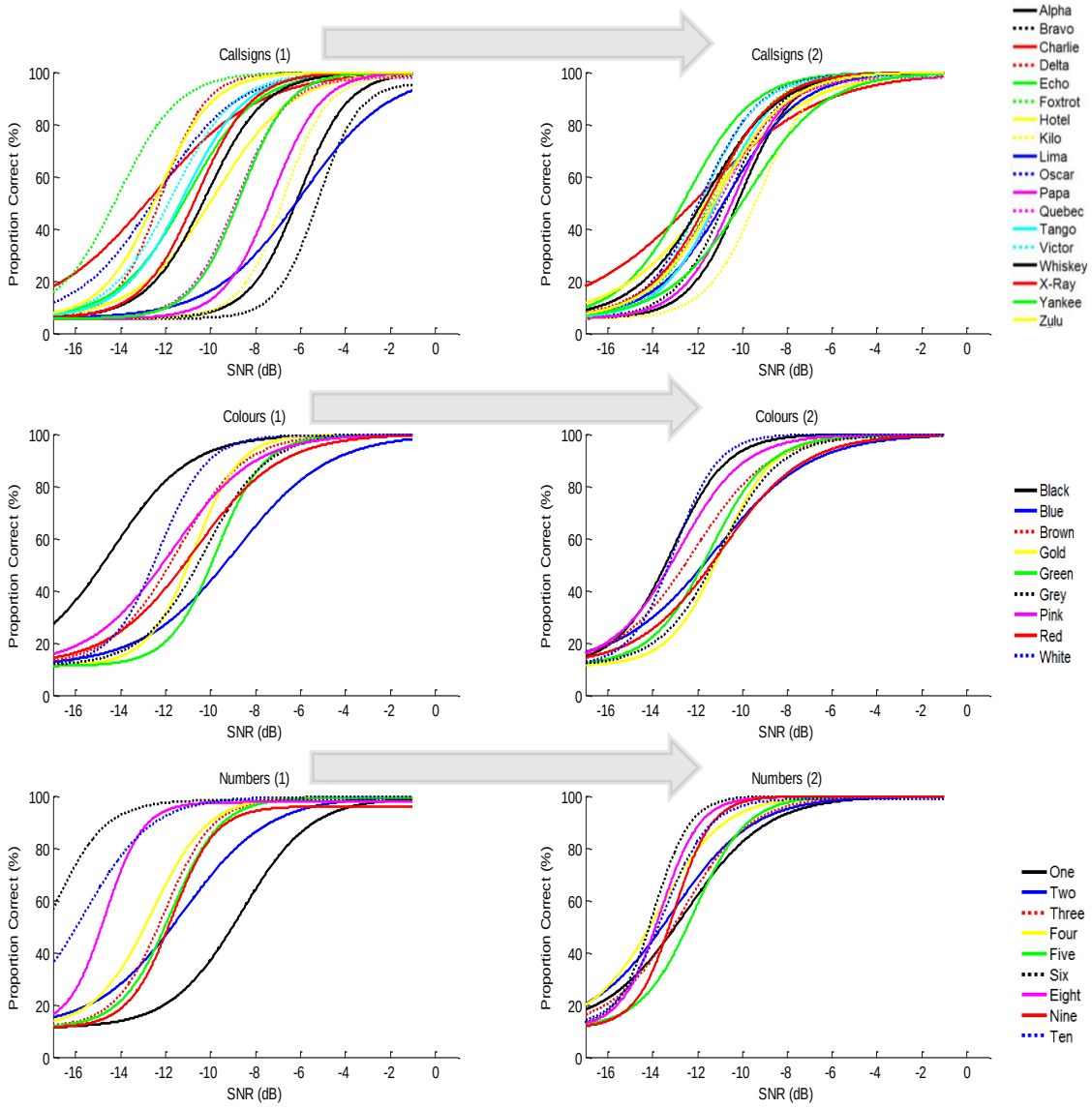


Figure A.4 *Comparison of PFs of the CRM target words before and after intelligibility equalisation*

Appendix H Graphs showing Bland and Altman plots for between session and repeat changes in SRT score

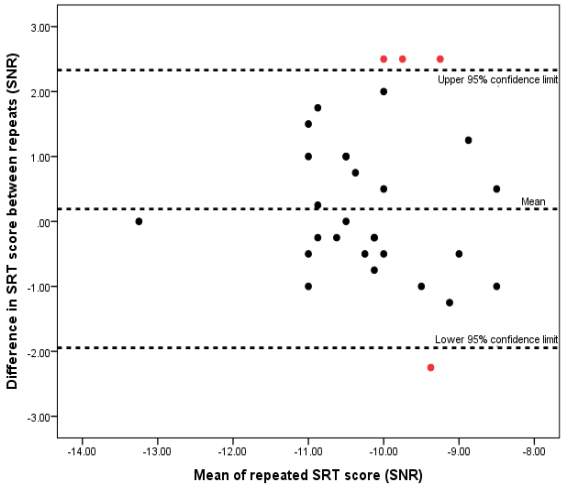
A set of Bland and Altman plots have been created separately for the normal hearing and hearing impaired data to look at the within subject variation between repeats and between sessions for the normal hearing listeners . For the normal hearing data three plots have been created for each test condition; two for the within session repeats and one for between sessions. The between session plots look at the differences between the first SRT measurement for each session. For the hearing impaired data there is only one plot for each test condition for the within session repeats.

The Bland and Altman plots show the difference between SRT across the repeats (always the first repeat minus the second, e.g. repeat three minus repeat four) on the *y* axis and the mean of the repeated SRT measurements plotted on the *x* axis. If there was no difference between the repeats, the mean difference (the central dotted line) would lie at zero on the *y* axis and all the points would lie along this line. If the variation between repeats changes across SRT scores then a trend in the difference would be seen. For example, a greater spread in the difference between repeats for the worse SRT scores would indicate poorer repeatability for this population. The upper and lower 95% confidence limits (the top and bottom dotted lines on each plot) show the 1.96 standard deviation of the mean differences.

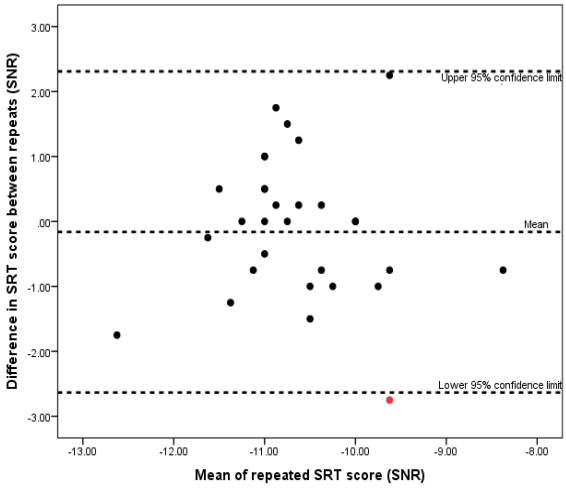
Appendix H

Normal Hearing

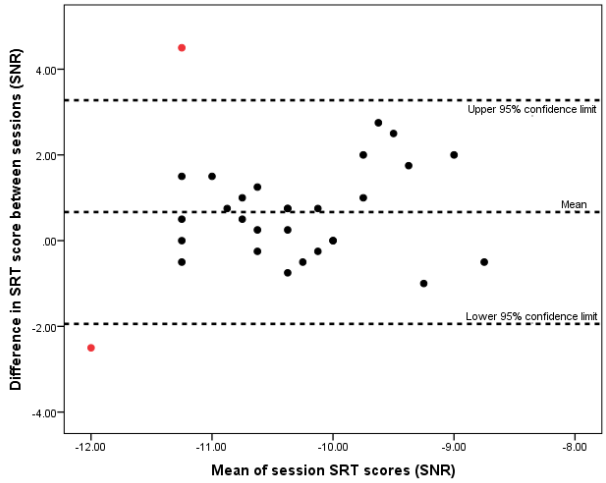
CRM-CSoff (repeats one and two)



CRM-CSoff (repeats three and four)

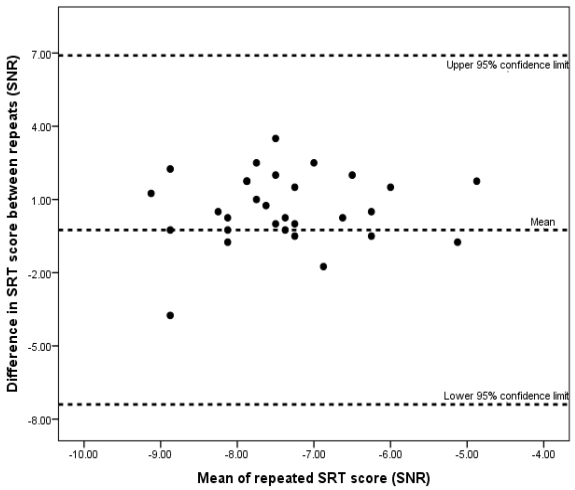


CRM-CSoff (sessions one and two, comparing repeats one and three)

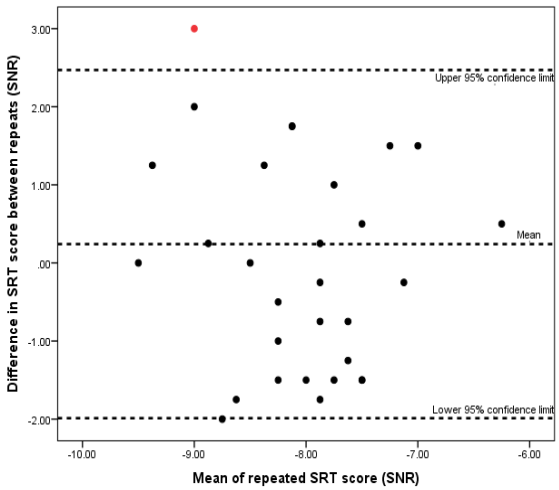


Normal Hearing continued

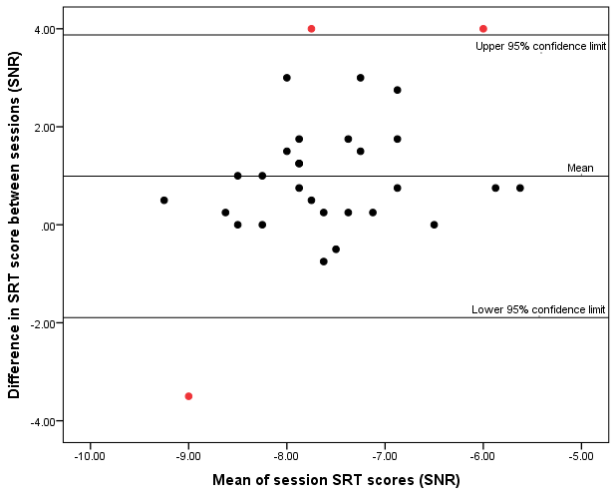
CRM-CSon (repeats one and two)



CRM-CSon (repeats three and four)

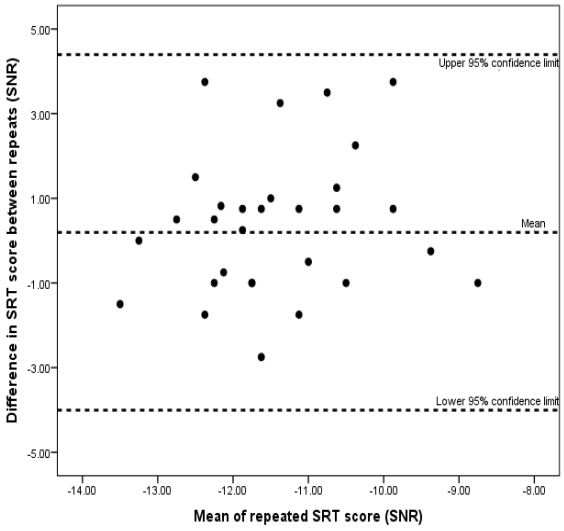


CRM-CSon (sessions one and two, comparing repeats one and three)

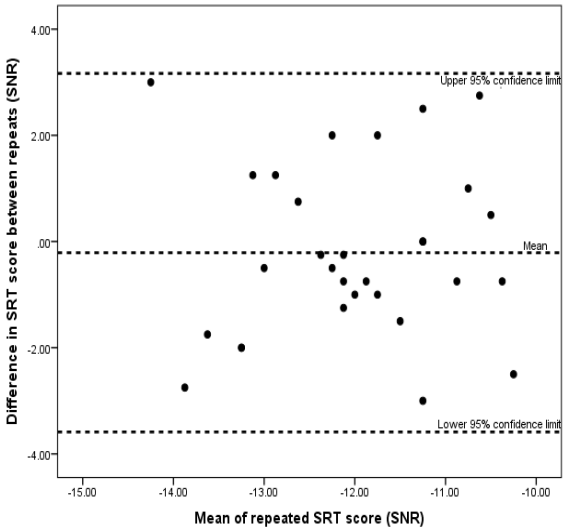


Normal Hearing continued

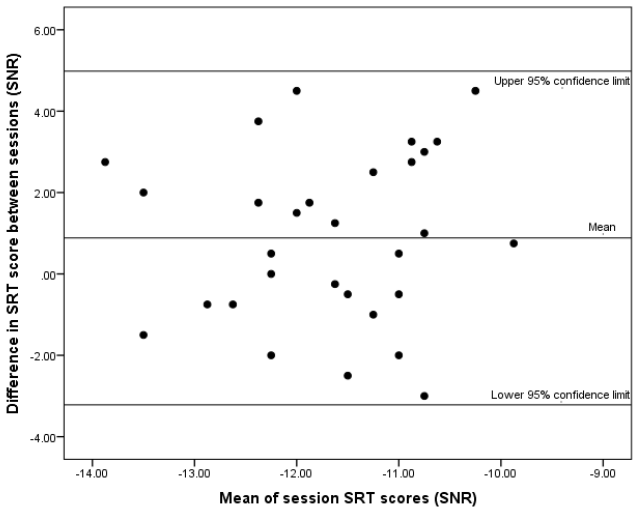
TDT (repeats one and two)



TDT (repeats three and four)

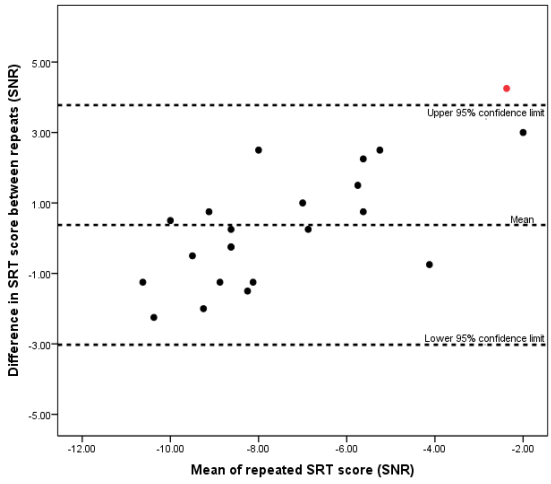


TDT (sessions one and two, comparing repeats one and three)

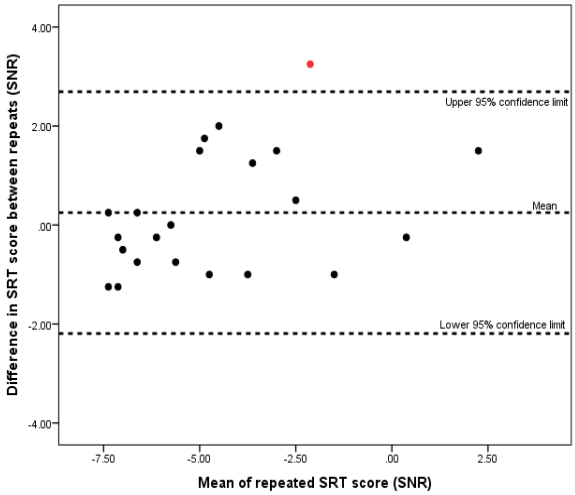


Hearing Impaired

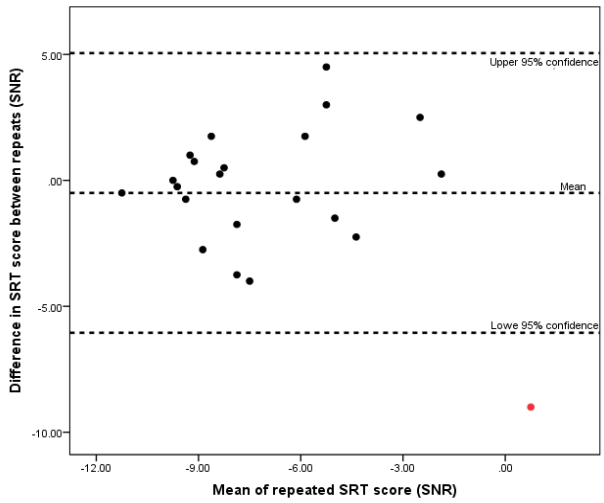
CRM-CSoff (Repeats one and two)



CRM-CSon (Repeats one and two)



TDT (Repeats one and two)



Appendix I Visual awareness tasks development- 'Tag the Enemy'

An Android Application has been created which was designed to simulate the some of the cognitive abilities required to carry out both an auditory and visual awareness task at the same time. The application was created by Laurence Charles, a postgraduate student at the department of Electronics and Computer Science, for the dissertation element of his master's degree. The thesis author designed the application and the coding was carried out by Laurence. The author had a supervisory role in collaboration with Gary Wills (Associate Professor in Computer Science) throughout Laurence's project. Laurence successfully passed his masters dissertation.

The aim of the project was to design a visual awareness task in which participants were required to respond to movement on a screen. This involved a 2D military scene in the background and an image of a soldier which moved to random location on the screen at randomised intervals. The participant was required to touch the screen to 'tag the enemy' as fast as possible as they appear on the screen. The administrator was able to alter the configuration of the game. The adjustable factors included: 1) the total number of enemy appearance in one game; 2) the length of time each appears for; 3) the minimum and maximum intervals between enemy appearances; 4) the total game duration; 5) cycle time for the background screen to change. The application also allows the administrator to export the results of each participant as a CSV file, providing information about the number of hits and misses and reaction times for each trial.

Due to the complexities of designing and running a dual-task experiment it was decided that this application would not be incorporated into the initial MCAT simulation development (see Section 6.2). The game has therefore not been trialled at this stage. However, details about the game, named 'Tag the enemy', are included below.

(1) Main Menu

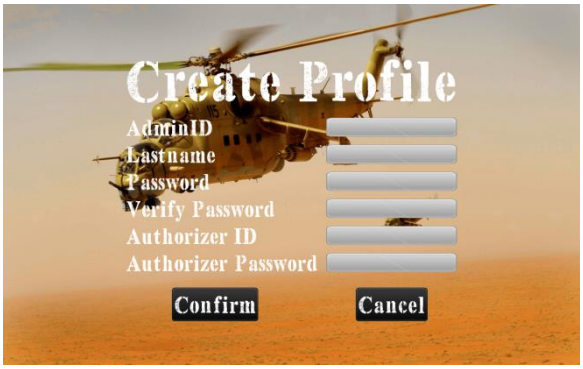
(2) Main Menu → Instructions

(3) Main Menu → Administrator

This page allows the administrator to either register as a new user or login as a previous user. Under 'Instructions' is the same screen as Main Menu → Instructions.

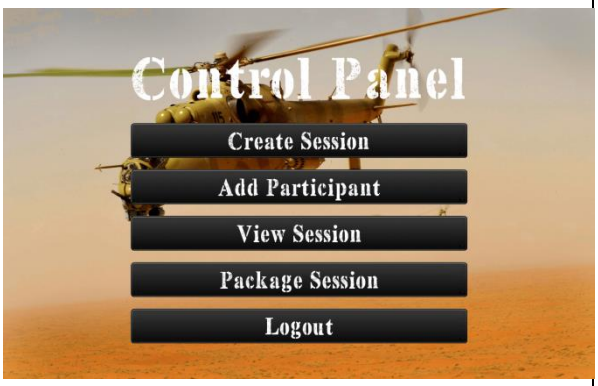
(4) Main Menu → Administrator → Register

Here a new user can enter their log in details and password.



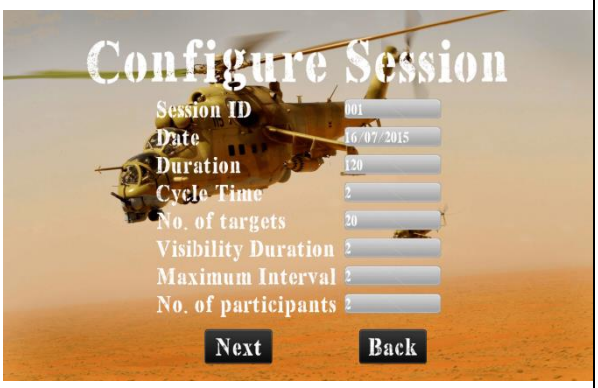
(5) Main Menu → Administrator → Log in

This screen appears after an administrator enters their correct log in details.



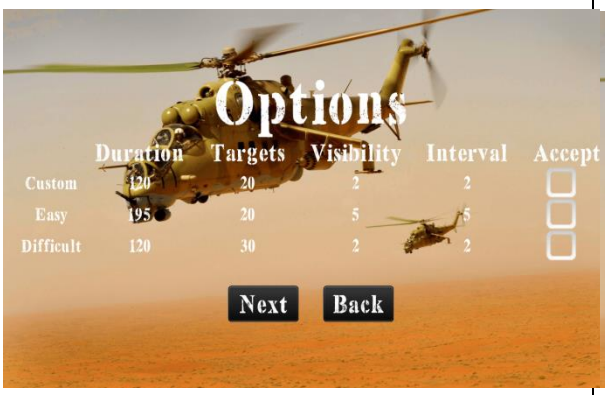
(6) Main Menu → Administrator → Log in → Create Session

Here the administrator is able to configure the game settings, adjusting each of the parameters listed.



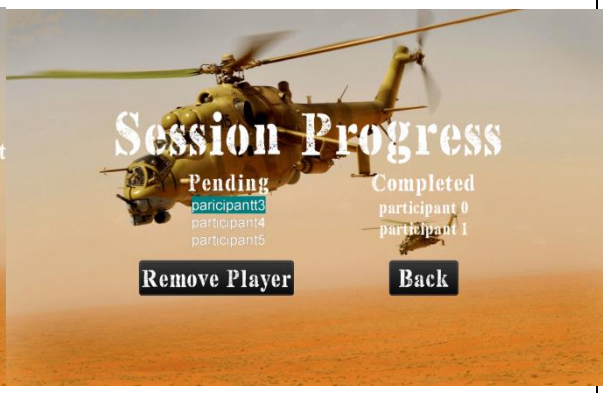
(7) Main Menu → Administrator → Log in → Create Session → Next

Based on the numbers entered on the previous screen the administrator is presented with 3 options for game play. Custom: using values entered on previous screen. Easy: targets displayed for maximum amount of time with maximum intervals. Hard: targets displayed for minimum amount of time with minimum intervals.



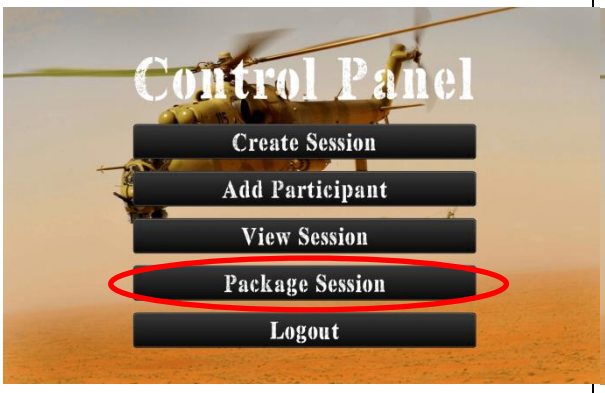
(8) Main Menu → Administrator → Log in → View Session

This screen allows the administrator to view how many participants have completed or still need to complete the current session.



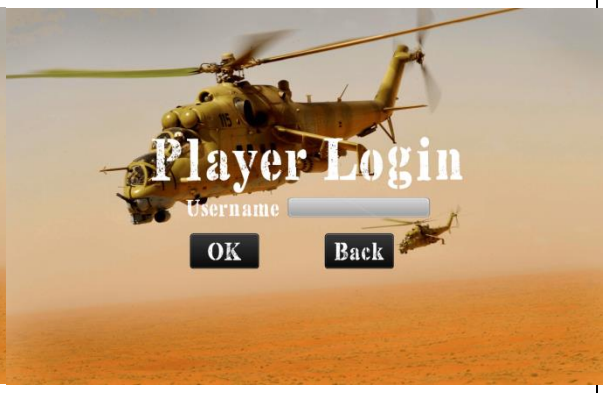
(9) Main Menu → Administrator → Log in → Package Session

By selecting this button the results of each participant are exported into CSV files and stored on the tablet. The file reports the number of hits and misses and reaction time. These can then be saved in Excel for analysis.



(10) Main Menu → Player

Here the player enters the username provided to them by the administrator.



(11) Main Menu → Player → Login

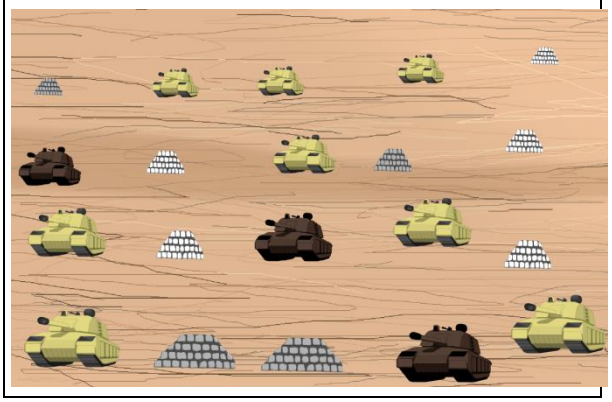
Here the participant has only two options, to start the game or read instructions about play (but these were not completed).

(12) Main Menu → Player → Login → Start Game

This is a still shot of game play image. The 'enemy' (middle left here) moves randomly, as does the background screen. The player is required to tap the enemy when they appear.

(13) Main Menu → Player → Login → Start Game

A second still shot of game play image. Here the 'enemy' is waiting to appear and the background screen has changed.



Appendix J List of recorded commands for each speech communication MCATs

TASK GROUP	COMMAND
T1: Accurately hearing commands in a casualty situation	1. MAN DOWN! MAN DOWN! 2. I REQUIRE THE **/MEDIC/**AT**/MY LOCATION/NOW/** 3. **/CASUALTY/**HAS **/GUN SHOT WOUND/**TO **/LEFT ARM/** 4. **/WE HAVE 3 /T1 CASUALTIES/** 5. GET THE CASUALTY ON THE STRETCHER AND TO THE HLS ASAP 6. 3 SECTION, FORM ALL ROUND DEFENCE IN THE VICINITY OF THE CASUALTY 7. 1 SECTION, CONFIRM YOU HAVE ALL YOUR MEN 8. GET AN FFD AND TOURNIQUET ON HIS RIGHT LEG, AS SOON AS POSSIBLE 9. WE HAVE A T-ONE CASUALTY AT MY LOCATION, 1 TIMES TRIPLE AMPUTEE 10. CAS REP BRAVO VICTOR 8974 GUN SHOT WOUND TO RIGHT ARM
T2: Accurately hearing grid references	1. **/CONFIRM/**YOU ARE NOW **/AT GRID 387489/** 2. **/MOVE NOW/**TO **/GRID 236 796/** 3. **/I AM NOW MOVING/**TO **/GRID 451 667/**TO **/MEET** THE **LIASON OFFICER/** 4. CALL SIGN 30 ALPHA'S VEHICLE HAS HIT AN IED AT GRID 517 444 5. **/I CAN CONFIRM/**THAT THERE ARE **/ENEMY FORCES/**AT **/GRID 917 048/** 6. **/ONE TIMES ENEMY FORCE/SEEN DIGGING IN AN IED/**AT **/GRID 403 557/** 7. **/I WANT YOUR SECTION/**TO **/FORM A HASTY DEFENCE/**AT **/GRID 202 312/** 8. **/ALL MY CALL SIGNS/HAVE NOW GONE FIRM/AT GRID 146 787/** 9. THE GRID FOR **/TONIGHT'S AMBUSH/**WILL BE **/758 671/** 10. YOUR SECTION NEED TO BE AT GRID 479 568 ASAP 11. **/YOUR CALL SIGN/**IS TO **/CONDUCT A VCP/AT GRID 602 706/**
T3: Accurately hearing directions on patrol	1. TAKE YOUR SECTION AND MOVE TO BUND LINE AND FORM A FIRE BASE 2. POSSIBLE ENEMY FORCES SEEN AT THE END OF THE TRACK 3. COME FORWARD AND COVER DOWN THE EDGE OF THE WOODLINE 4. TWO ONE CHARLIE MOVE EAST AND COVER THE HIGH GROUND 5. YOUR SECTION WILL MOVE TO THE COMPOUND FIRST 6. KEEP YOUR SPACING AND ENSURE THE REAR MAN COVERS THE REAR 7. MOVE INTO THE WOOD LINE AND GO FIRM 8. GET EYES ON THE TARGET AREA NOW 9. GET ONE OF YOUR MEN TO SPEAK WITH THE CIVILIAN ON THE TRACK 10. 2IC MOVE LEFT AND COVER THE HIGH GROUND 11. I REQUIRE AN ACCURATE DESCRIPTION OF THE WEAPONS CARRIED BY THE ENEMY
T4: Accurately hearing directions in a vehicle	1. **/AT THE FORK/IN THE TRACK/GO LEFT/** 2. **/STOP/**AT THE **/EDGE OF THE BUILDING/** 3. **/SLOW DOWN/**AND **/KEEP YOUR DISTANCE/** 4. **/ENSURE/YOU KEEP IN/THE DEAD GROUND/** 5. **/PICK UP THE PACE/** 6. **/MOVE EAST/**AND **/COVER HIGH GROUND/** 7. **/PREPARE TO MOVE/** 8. WE WILL DEPLOY LEFT….YOU WILL DEPLOY RIGHT, OK? 9. **/LIGHTS OFF NOW/!!** 10. WE WILL MOVE IN TWO HUNDRED METRE BOUNDS ON BEARING ZERO NINE HUNDRED MILS 11. **/WE WILL CONDUCT/**A **/SHORT HALT/**AT **/THIS LOCATION/** 12. I WANT YOUR TOP COVERS TO ADOPT ALTERNATE ARCS

T5: Accurately hearing fire control orders	1. **/ENEMY/FORCES SEEN/**AT THE **/EDGE OF THE BUILDING/**
	2. CHARLIE FIRE TEAM, TWO TIMES ENEMY, EDGE OF WOODLINE, RAPID…… FIRE
	3. TWO SECTION, FOUR TIMES ENEMY, TWO HUNDRED METRES TO YOUR FRONT, RAPID…….. FIRE
	4. TWO SECTION, ONE HUNDRED METRES, SMALL COPSE, BOTTOM BASE OF COPSE, ENEMY
	5. **/GUNNER/GET SOME SUPPRESSIVE FIRE/ONTO THAT POSITION/**
	6. GRENADIER, MOVE FORWARD AND TAKE OUT THE BUNKER
	7. SECTION, ENEMY FORCES IN THE HEDGE LINE, WATCH MY TRACER
	8. **/SECTION/, /TWO TIMES ENEMY FORCES/IN TOP WINDOW/, AWAIT MY ORDER/**
	9. MACHINE GUNNER, ENEMY TO YOUR FRONT, IN BURSTS, RAPID……FIRE
	10. DELTA FIRE TEAM, THREE TIMES ENEMY FORCES IN BUNKER, RAPID……FIRE
T6: Accurately hearing 'stop' commands	**1. /ALL CALLSIGNS/GO FIRM/**
	2. STOP STOP STOP!
	3. **/MOVE INTO/THE DITCH/**AND **/GO FIRM/**
	4. **/GO FIRM/**AND **/AWAIT MY ORDERS/**
	5. /HALT!!/ADVANCE AND BE RECOGNISED/
	6. **/HOLD IT THERE/!**
	7. **/THREE ONE CHARLIE/GO FIRM/ON THE TRACK/**AND **/OBSERVE TO THE SOUTH/**
	8. TWO ONE DELTA STOP THERE! POSSIBLE ENEMY FORCES SEEN IN BUILDING
	9. DO NOT MOVE FORWARD OF THE BUND LINE
	10. ALL CALL SIGNS GET DOWN ON YOUR BELT BUCKLES
T7: Accurately hearing the briefing before a foot patrol	1. TONIGHT WE WILL BE CONDUCTING A STANDING PATROL
	2. THE ENEMY FORCES MORALE IS LOW DUE TO RECENT ATTRITION RATES
	3. ONE SECTION POINT SECTION, TWO SECTION WILL FORM A FIRE BASE, THREE SECTION RESERVE
	4. THE GROUND IS OPEN WITH VERY LITTLE DEAD GROUND TO USE
	5. THE CSM WILL DEAL WITH ALL CASUALTIES
	6. THREE SECTION WILL CONDUCT A RECCE PATROL AT CROSSING POINT CHARLIE
	7. ALL CASUALTIES WILL BE LAID ON THE AXIS
	8. ONE SECTION ARE TO MOVE THE EDGE OF THE WOODLINE TO THE NORTH
	9. ALL CALL SIGNS ARE TO RE-ORG AT GRID 613 782
	10. TWO SECTION WILL FORM A FIRE BASE
	11. YOUR MISSION IS TO CLEAR ALL ENEMY IN BOUNDARIES NO LATER THAN TWELVE HUNDRED HOURS

Appendix K Pilot Experiment: selecting an SNR for the VEHCOM SimMCAT

In Study 6 (Section 6.4) it was necessary to select a single SNR to run the VEHCOM SimMCAT at (see Section 6.3 for the development of this test). The test could only be run at one SNR because it was not possible to present the stimuli more than once at different SNRs without avoiding learning effects. There are two main considerations when selecting an SNR: 1) representative of the levels when listening over a radio in a vehicle and 2) avoiding floor and ceiling effects.

Selecting a single 'realistic SNR' is problematic since it is dependent on numerous variables such as the speed the vehicle is travelling at, the terrain and the volume the radio is set at. In addition, because of issues relating to safe levels of noise exposure within experiments it would not be possible to replicate the exact conditions of this scenario for ethical reasons. Instead, selecting an SNR which would avoid floor and ceiling effects across the hearing acuity groups was prioritised. It was important that the chosen SNR allows for those assigned to the severe hearing acuity group to hear some of the commands (not always scoring 0%) and for the normal hearing group to not be consistently scoring 100% correct.

In order to obtain a general idea of performance levels, 0 dB SNR was selected as the initial starting point and six participants took part in the pilot experiment (normal n=2, moderate n=2 and severe n=2). In order to minimise testing time only one of the intermediate hearing acuity groups (moderate) was piloted, since the focus was on whether the normal and severe hearing acuity groups avoided floor and ceiling effects. As shown in Figure A.5 (black columns) at 0dB SNR the normal hearing listeners were scoring highly (86%) and the severe hearing acuity group were also scoring fairly highly (33%). It was decided that the SNR could be lowered to decrease the performance of the two groups and potentially be a more representative of the difficult listening environment experienced in a real world scenario listening over a radio in an armoured vehicle. A second pilot was run with stimuli presented at -5 dB SNR and the results are also shown in Figure A.5 (white columns). Eight civilians participated in the second pilot (n=2 for each hearing acuity group). The scores did not differ a great deal from those obtained at 0dB SNR for the normal and moderate hearing acuity groups but performance for the severe group decreased by roughly 15%. It was thought that a further SNR reduction may cause the severe hearing acuity group to reach floor effects; for this reason no further SNRs were tested and -5dB SNR was chosen. This was also

not thought to be such an advantageous SNR that it is unrealistic of a real world listening environment for the given scenario.

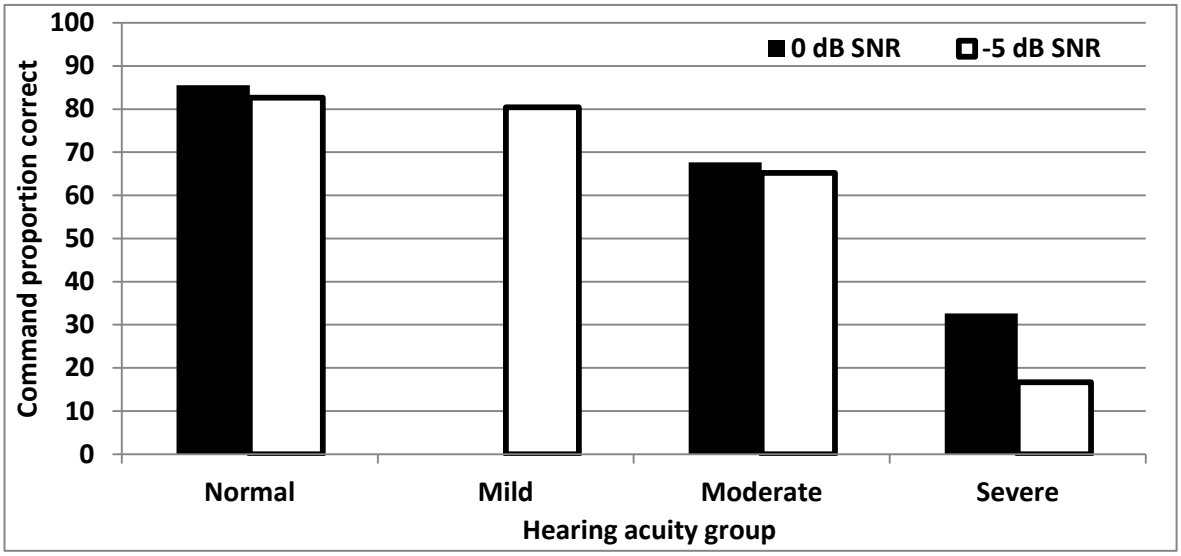


Figure A.5 *Pilot experiment results. Comparing performance levels listening to commands at 0 and -5 dB SNR, across four levels of hearing acuity.*

Appendix L List of MATLAB Code Authors

MATLAB Code and Study	Author contribution to code	Main author
CRM method of constant stimuli (Study 3)	Wrote pseudocode and typed MATLAB code with instruction from Carla Hampson.	Carla Hampson (acquaintance and Operations Analyst at Defence Science and Technology Laboratory)
CRM adaptive procedure (Study 4)	Wrote pseudocode	Daniel Rowan (PhD Supervisor)
Hearing loss simulation software (Study 6)	None (existed prior to this research)	Jessica Monaghan (Research Fellow, University of Southampton)
CRM adaptive procedure including hearing loss simulation (Study 6)	Wrote pseudocode to instruct code author on how the hearing loss simulator should be incorporated into the CRM adaptive procedure code	Daniel Rowan (original CRM adaptive procedure code) and Jessica Monaghan (modified CRM adaptive procedure code including hearing loss simulation)
VEHCOM SimMCAT (Study 5 & 6)	Wrote detailed pseudocode and some aspects of the MATLAB code. Worked alongside Falk-Martin to complete the MATLAB Code.	Falk-Martin Hoffman (Postgraduate Researcher, University of Southampton)

Appendix M Confirmation of ethics approval

Study 1 part A: exploring auditory tasks

MOD Research Ethics Approval

MOD Research Ethics Committee (General)
Corporate Secretariat
Bldg 5, G01-614
Dstl Porton Down
Salisbury, Wiltshire
SP4 0JQ
Secretary: Marie Jones
telephone: 01980 658155
e-mail: mnjones@dstl.gov.uk
fax: 01980 613004

Miss Zoe Bevis
Hearing and Balance Centre
Institute of Sound and Vibration Research
Building 13
University of Southampton
Southampton
SO17 1BJ

Ref: 359/GEN/12

Dear Miss Bevis,

Re: Identification of key listening situations for military personnel – 1st amendment

I am happy to give ethical approval for this amendment to allow Gülcan Garip to be added to the list of investigators.

You also enquire about gathering further information about the tasks through discussion with subject matter experts. This does not need further ethical review.

Yours sincerely,

Robert Linton

Dr Robert Linton
Chairman MOD Research Ethics Committee (General)

telephone: 020 8877 9329
e-mail: robert@foxlinton.org
mobile: 07764616756

University of Southampton Ethics Approval

From: ERGO [mailto:ergo@soton.ac.uk]
Sent: 07 May 2013 12:57
To: Shutt H.D.
Subject: Your Ethics Submission (Ethics ID:5850) has been reviewed and approved

Submission Number: 5850
Submission Name: Identification of mission-critical hearing tasks for infantry personnel
This is email is to let you know your submission was approved by the Ethics Committee.

Comments
None

Study 1 part B: identification of MCATs

MOD Research Ethics Approval

From: NAVY INM-EMS HEAD (Allsopp, Adrian B1) <NAVYINM-EMSHEAD@mod.uk>
Sent: 02 July 2013 08:37
To: Shutt H.D.
Subject: MCAT Questionnaire

Hannah,

I'm content that we can justify your MCAT questionnaire as "further information about tasks" and this is covered by the present ethical approval dated 13 Mar 13.

Adrian.

Dr Adrian J. Allsopp PhD, MSc, MPA, PGCE, BSc (Hons) **Head of Environmental Medicine and Science, Institute of Naval Medicine, Crescent Road, Alverstoke, Hampshire PO12 2DL** ☎Mil: **9360 68066**Civ: **+44(0)2392 768066** Mob: **0783 7217461**DII(F)
NAVY INM-EMS HEAD|✉ Email: **NAVYINM-EMSHEAD@mod.uk** | **Adrian.Allsopp741@mod.uk**

To promote, protect and restore the health of the Royal Naval Service

University of Southampton Ethics Approval

From: ERGO [mailto:ergo@soton.ac.uk]
Sent: 06 August 2013 12:59
To: Shutt H.D.
Subject: Your Ethics Submission (Ethics ID:6686) has been reviewed and approved

Submission Number: 6886
Submission Name: Questionnaire: The Identification of Mission-Critical Hearing Tasks for Infantry Personnel
This is email is to let you know your submission was approved by the Ethics Committee.

Comments
1.Thanks for clear and helpful responses to our queries as presented in the separate document, and also for highlighting the changes in other documents. All is OK ... but please check that you have uploaded the correct version of the questionnaire ... the one with correct logos.

Study 2: Developing and recording CRM speech-in-noise test

No ethics required as no data collection carried out

Study 3: Equalising the intelligibility of the CRM in noise

MOD Research Ethics Approval

Not required, not military personnel participated in this study

University of Southampton Ethics Approval

From: ERGO [mailto:ergo@soton.ac.uk]
Sent: 03 April 2014 12:29
To: Shutt H.D.
Subject: Your Ethics Submission (Ethics ID:9762) has been reviewed and approved

Submission Number 9762:
This email is to confirm that the amendment request to your ethics form (Measuring the internal validity of the coordinate response measure speech intelligibility in noise test (Amendment 1))has been approved by the Ethics Committee.

You can begin your research unless you are still awaiting specific Health and Safety approval (e.g. for a Genetic or Biological Materials Risk Assessment)

Comments
None

Study 4: Evaluating the measurement precision of the CRM implemented in an adaptive procedure

MOD Research Ethics Approval

Ministry of Defence
Research Ethics Committee

From the Chairman
Professor Allister Vale
National Poisons Information Service (Birmingham Unit),
City Hospital, Birmingham B18 7QH

Telephone: 0121 507 4123
e-mail:allistervale@npis.org

Mrs Hannah Shutt
Postgraduate Researcher and Audiologist
University of Southampton
Hearing and Balance Centre
ISVR, Building 13
University of Southampton
S017 1BJ

Our Reference:
505/MODREC/14

Date: 31 March 2014

Dear Mrs Shutt,

Thank you for submitting your revised Protocol 505 with tracked changes, and with a covering letter with responses to my own letter. The revised protocol has been approved by the Officers of MODREC ex-Committee.

I wish you and your colleagues a successful study and we look forward to receiving in due course a brief summary of the results so that these can be filed in accordance with the arrangements under which MODREC operates.

Yours sincerely

Allister Vale MD FRCP FRCPE FRCPG FFOM FAACT FBTS FBPhamacolS Hon FRCPSG

cc Professor David Jones, Dr Paul Rice OBE, Marie Jones

University of Southampton Ethics Approval

From: ERGO [mailto:ergo@soton.ac.uk]
Sent: 11 February 2015 10:34
To: Shutt H.D.
Subject: Your Ethics Submission (Ethics ID:13712) has been reviewed and approved

Submission Number 13712:
This email is to confirm that the amendment request to your ethics form (Measuring the internal validity of the coordinate response measure speech intelligibility in noise test (Amendment 2)) has been approved by

the Ethics Committee.

You can begin your research unless you are still awaiting specific Health and Safety approval (e.g. for a Genetic or Biological Materials Risk Assessment)

Comments
1.I am giving this study approval however there a couple of points that I feel ought to be dealt with before the study starts. Participant Information sheet (PIS) - Martina Prude is no longer the contact point for persons to raise concerns, nor is her email address valid. Complaints and concerns should be addressed to the "Research Governance Manager", RGO will be able to supply appropriate contact details. Reward for taking part - the study advert mentions Â£20 but does not indicate whether this is per session or for all 2-3 sessions. Also the reward is not mentioned in the PIS. Both advert and PIS should cover any rewards being offered and be specific

Study 5: Developing and recording the VEHCOM SimMCAT

No ethics required as no data collection carried out

Study 6: Measuring the predictive validity of the CRM as a measure of military AFFD and exploring the importance of experience when assessing AFFD

MOD Research Ethics Approval

Ministry of Defence
Research Ethics Committee

From the Chairman
Professor Allister Vale MD
National Poisons Information Service (Birmingham Unit),
City Hospital, Birmingham B18 7QH

Telephone: 0121 507 4123
e-mail:allistervale@npis.org

Mrs Hannah Shutt
Postgraduate Researcher and Audiologist
Institute of Sound and Vibration Research (ISVR)
University of Southampton
Hearing and Balance Centre, ISVR, Building 13,
University of Southampton, SO17 1BJ

Our Reference:
584/MODREC/14

Date: 30 November 2014

Dear Mrs Shutt,

The development of an improved hearing test for military personnel- introducing the Coordinate Response Measure speech-in-noise test

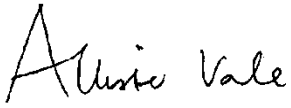
Thank you for submitting your revised Protocol 584 with tracked changes, and with a covering letter with responses to my own letter. The revised protocol has been approved by the Officers of MODREC ex-Committee with correction of a small number of typographical errors, including the name of MODREC!

I wish you and your colleagues a successful study. In due course please send the Secretariat a

final report containing a summary of the results so that these can be filed in accordance with the arrangements under which MODREC operates**.** Please would you also send a brief interim report in one year's time if the study is still ongoing.

This approval is conditional upon adherence to the protocol – please let me know if any amendment becomes necessary.

Yours sincerely



Allister Vale MD FRCP FRCPE FRCPG FFOM FAACT FBTS FBPhamacolS FEAPCCT Hon FRCPSG

cc, Professor David Jones, Professor David Baldwin, Marie Jones

University of Southampton Ethics Approval

From: ERGO [mailto:ergo@soton.ac.uk]
Sent: 02 December 2014 14:12
To: Shutt H.D.
Subject: Your Ethics Submission (Ethics ID:12949) has been reviewed and approved

Submission Number: 12949
Submission Name: The development of an improved hearing test for military personnel- external validation of the coordinate response measure speech-in-noise test
This is email is to let you know your submission was approved by the Ethics Committee.

You can begin your research unless you are still awaiting specific Health and Safety approval (e.g. for a Genetic or Biological Materials Risk Assessment)

Comments
None

Appendix N List of presentations

June 2015: Imperial College Ear-Monitoring Workshop

Representing the Hear for Duty team, I presented a poster providing an overview of our team's work. I contributed towards a workshop discussing the potential introduction of an in-ear sensing device for military personnel. As the only audiologist at the workshop I provided valuable expertise about auditory issues relating to the product design and wear-ability.

March 2015: Royal Centre for Defence Medicine

The Hear for Duty PhD students were invited by Brigadier Tim Hodgetts, the Medical Director at Defence Medical Services to present our research work and to discuss future work in our field. A team of researchers from the US Department of Defence Hearing Centre of Excellence also attended to find out about our work and to report their progress in a similar field. The success of this meeting contributed towards the Hear for Duty team being encouraged to submit six funding proposals to Brigadier Hodgetts in September '15.

November 2014: British Academy of Audiology Annual Conference Presentation

After submitting an abstract to the conference I was invited to present my PhD research findings as an aural presentation to a mixed audience of academics and clinicians.

August 2014: Army Hearing Working Group

I was invited to attend this meeting specifically to update members of the working group about the research activity occurring at Southampton University. I delivered a short presentation with an accompanying hand-out which provided an overview of our research focus with the aim of communicating the importance of the group continuing to support work in this field.

May 2014: Three Minute Thesis Competition

Winner of the intra-faculty competition and second prize in the University Final.

February 2013: Institute of Naval Medicine Journal Club-

Following the completion of the first experiment of my PhD (published in Noise and Health, Bevis et al. 2014, see publication list) I co-presented the findings to a non-specialist audience at the Institute of Naval Medicine Journal Club

Appendix O Participant recruitment challenges and strategies in the military population

When conducting research on human participants, and in particular when recruiting from specific populations, recruiting enough participants and completing experiments to time can be a challenge. The military population is no such exception. As a result of the work conducted in this thesis (in particular studies one, four and six) the author can provide a novel insight into the specific recruitment challenges encountered when working with the UK Ministry of Defence and provide some strategies for overcoming these challenges.

Challenge 1: Military personnel follow a tight training schedule and have very little free time whilst on base. In addition, personnel are understandably reluctant to give up this free time. This was a particular challenge for Study 1. Three responses from senior personnel stated that they would not be able to find a suitable time to distribute and complete the questionnaire within the timeframe of the study due to training or deployment commitments.

Suggested strategy 1: The researcher must be extremely flexible with regards to data collection times and also be prepared to extend the time frame of the study in order to collect adequate data. Contacting those who will help you run the study a long time in advance (up to six months) to ask them to block out some time for the study to run is helpful. Ensuring the senior officers who are facilitating the running of the study are aware of the importance of the work helps with encouraging them to find time for data collection.

Challenge 2: In Study 4 the author was recruiting personnel from the list of patients at the Defence Audiology Service (DAS) in Gosport. This is only DAS centre in the UK so patients travel from all over the UK to attend appointments there. It was therefore only possible to ask patients to participate if they were already travelling to DAS. This meant recruitment was limited to patients who met the participation criteria, had appointments during the timeframe of the study and were willing to take part and were not in a rush to travel home. The author had to be very flexible, running the experiment on any days that patients were free and often travelling to Gosport to only test one participant or to find out that a potential participant had not arrived for their appointment.

Suggested strategy 2: Again, the researcher must be extremely flexible as to when they are able to travel to collect data. Using a recruitment method which allows the researcher to contact the potential participant on the day of testing to confirm their appointment attendance would have

been beneficial. If the experimental budget allows then inviting individuals to a central study centre or funding participants to travel to DAS just to participate would increase the available population and speed up recruitment and data collection.

Challenge 3: Ensuring that the recruitment procedure follows ethical boundaries can be challenging in the military environment. The hierarchical culture within the military means that senior personnel instruct those under their command to carry out tasks and less senior personnel do not routinely question what is being asked of them. It is therefore challenging to ensure that participants understand that their participation is entirely voluntary and they can withdraw at any time without giving reason. This was a particular challenge for studies one and six.

Suggested strategy 3: The researcher spent time with the senior personnel who helped with the running of the studies to explain the importance of participants being aware that the study was entirely voluntary. Before starting the studies the researcher reiterated this point to participants and requested that they re-read the participant information sheet prior to signing the consent form.

Challenge 4: Obtaining a network of MoD contacts that are willing to help with the logistics and running of experiments can be challenging as a civilian.

Suggested strategy 4: Approaching institutes or companies who have previous experience working with the MoD and asking them to put you in touch with relevant personnel is helpful. The author liaised with the Institute of Naval Medicine to obtain contacts and attended the Army Hearing Working Group who provided a contact to help run Study 6. When approaching military personnel it is important to very briefly summarise the key aims of the study, focussing on the benefits to the MoD, and explicitly detail what help you are requesting. Once the author obtained a network of contacts they found they were extremely helpful.

Challenge 5: Obtaining ethics approval from the Ministry of Defence Research and Ethics Committee (MODREC) took a long time (4-6 months) and therefore influenced the start date of experiments.

Suggested strategy: A great deal of forward planning is required in order to obtain ethical approval without delaying the start of data collection. Completing the application in collaboration with someone with previous experience is very helpful. The author has since been informally told that the maximum amount of time an ethics application will take will be 2-4 months. An awareness of challenge 3 should influence suggested recruitment strategies.

References

Akeroyd, M.A., 2008. Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults. *International Journal of Audiology,* 47 (Suppl. 2), p.S53-71.

Altman, D.G., Machin, D., Bryant, T.N. & Gardner, M.J., 2000. *Statistics with confidence*. 2nd ed. Bristol: British Medical Journal Books.

Alvord, L.S., 1983. Cochlear dysfunction in “normal hearing” patients with history of noise exposure. *Ear and Hearing*, 4, p.247-50.

Anastasi, A., 1988. *Psychological Testing*. 6th ed. New York: Macmillan.

Argyros, G.J., 1997. Management of blast injury. *Toxicology*, 121, p.105–15.

Baer, T., & Moore, B. C. J., 1993. Effects of spectral smearing on the intelligibility of sentences in noise. *The Journal of the Acoustical Society of America*, 94(3), p.1229–1241.

Baken, R.J., 1987. *Clinical Measurement of Speech and Voice.* London: Taylor and Francis Ltd.

Bell, S.A., 2001. *A beginner’s guide to uncertainty of measurement*. [online] National Physics Laboratory. Available at: <http://www.npl.co.uk/publications/a-beginners-guide-to-uncertainty-in-measurement> [Accessed 24 August 2015].

Bench, J., Kowal, A. & Bamford, J., 1979. The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children. *British Journal of Audiology*, 13(3), p.108–12.

Berman, G. & Rutherford, T., 2014. *Defence personnel statistics*. [online] House of Commons Library. Available at: <http://researchbriefings.parliament.uk/ResearchBriefing/Summary/SN02183> [Accessed 24 August 2015].

Bernstein, J. G. W., & Brungart, D. S., 2011. Effects of spectral smearing and temporal fine-structure distortion on the fluctuating-masker benefit for speech at a fixed signal-to-noise ratio. *The Journal of the Acoustical Society of America*, 130(1), p.473–488.

Bevis, Z.L., Semeraro, H.D.S., Rowan, D., van Besouw, R. & Allsopp, A., 2014. Fit for the frontline? A focus group exploration of auditory tasks carried out by infantry and combat support personnel. *Noise and Health*, 16(69), p.127–35.

Biggs, T., & Everest, A., 2011. British military hearing conservation programme. *Clinical Otolaryngology*, 36(3), p.299–301.

Bland, M., 2000. *An introduction to medical statistics.* 3rd ed. Oxford: Oxford University Press.

Bolia, R.S., Nelson, W.T., Ericson, M.A. & Simpson, B.D., 2000. A speech corpus for multitalker communications research. *Journal of the Acoustical Society of America*, 107(2), p.1065–6.

Boone, H.N. & Boone, D.A., 2012. Analyzing Likert Data. *Journal of Extension*, 50(2), Article Number 2TOT2.

Boothroyd, A., 1968. Developments in speech audiometry. *British Journal of Audiology*, 2, p.3-10.

Borg, E., Nilsson, R. & Liden, G., 1982. Fatigability of the Stapedius Reflex in Industrial Noise. *Acta Oto-laryngologica*, 94, p.385–93.

References

Bowers, L., Huisingh, R., & LoGiudice, C. 2006. *The Listening Comprehension Test 2*. East Moline: LinguiSystems, Inc. Available at: <https://www.linguisystems.com/products/product/display?itemid=10398>. [Accessed 30 October 2015].

Bowling, A., 2005. Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health (Oxford)*, 27(3), p.281–91.

Brand, T. & Kollmeier, B., 2002. Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *Journal of the Acoustical Society of America*, 111(6), p.2801–810.

British Army, 2011. *Speaking and Listening*. [online] Army Training and Education Documents. Available at: <http://www.army.mod.uk/training_education/24473.aspx> [Accessed 25 August 2015].

British Army, n.d. *British Army Vehicles and Equipment.* [online] Army Equipment Documents. Available at: <http://www.army.mod.uk/documents/general/285986_ARMY_VEHICLESEQUIPMENT_V12.PDF_web.pdf> [Accessed 26 August 2015].

British Library, 2014. *Received Pronunciation*. [online] British Library Language and Literature. Available at: <http://www.bl.uk/learning/langlit/sounds/case-studies/received-pronunciation/> [Accessed 25 August 2015].

British Society of Audiology, 2012. *Recommended procedure: Pure-tone air and bone conduction threshold audiometry with and without masking.* [online] British Society of Audiology Resources. Available at: <http://www.thebsa.org.uk/wp-content/uploads/2014/04/BSA_RP_PTA_FINAL_24Sept11_MinorAmend06Feb12.pdf> [Accessed 25 August 2015].

Brown, P.E.H. & Fallowfield, J.L., 2012. *The design and validation of a strength-based Royal Navy fitness test: Part 1- Identification of the most critically demanding generic tasks performed onboard Royal Navy ships.* [internal report] Institute of Naval Medicine, Gosport.

Brungart, D., Sheffield, B. & Grantham, M., 2013. *Evaluating the Operational Impact of Hearing Impairment.* [pdf] National Hearing Conservation Association. Available at: <http://c.ymcdn.com/sites/www.hearingconservation.org/resource/resmgr/imported/Brungart_DouglasNHCA2013v5.pdf> [Accessed 25 August 2015].

Brungart, D.S., 2001a. Evaluation of speech intelligibility with the coordinate response measure. *Journal of the Acoustical Society of America, 109(5), p*.2276–279.

Brungart, D.S., 2001b. Informational and energetic masking effects in the perception of two simultaneous talkers. *Journal of the Acoustical Society of America*, 110(5), p.2527–538.

Cain, P.A., 1998. Update - Noise Induced Hearing Loss and the Military Environment. *Journal of the Royal Army Medical Corps,* 144(2), p.97–101.

Campbell, D.T. & Stanley, J.C., 1966. *Experimental and Quasi-Experimental Designs for Research*, Chicago: Rand McNally.

Christiansen, C. & Dau, T., 2012. Relationship between masking release in fluctuating maskers and speech reception thresholds in stationary noise. *Journal of the Acoustical Society of America*, 132(3), p.1655–666.

Cohen, J., 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, p.37–46.

Coles, R.R.A., Lutman, M.E. & Buffin, J.T., 2000. Guidelines on the diagnosis of noise- induced hearing loss for medicolegal purposes. *Clinical Otolaryngology and Allied Sciences*, 25(4), p.264–73.

Cone, B., Dorn, P., Konrad-Martin, D., Lister, J., Ortiz, C. & Schairer, K., n.d. *Ototoxic Medications (Medication Effects)*. [online] American Speech-Language Hearing Association Public Resources. Available at: <http://www.asha.org/public/hearing/Ototoxic-Medications/> [Accessed 25 August 2015].

Cox, R.M., Alexander, G.C. & Gilmore, C., 1987. Development of the Connected Speech Test (CST). *Ear and Hearing*, 8(5 Suppl), p.119S-26S.

Crombie, I.K. & Davies, H.T., 2009. *What is a meta-analysis?* [online] What is…? Series. Available at: <http://www.medicine.ox.ac.uk/bandolier/painres/download/whatis/meta-an.pdf> [Accessed 13 November 2015].

Davis, M.H. & Johnsrude, I.S. 2007. Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research,* 229(1-2), p.132-47.

De Andrade, K. C. L., de Lemos Menezes, P., Carnaúba, A. T. L., de Sousa Rodrigues, R. G., de Carvalho Leal, M. & Pereira, L. D., 2013. Non-flat audiograms in sensorineural hearing loss and speech perception. *Clinics*, 68(6), p.815–19.

Defence Committee, 2003. *Recruitment.* [online] Select Committee on Defence Third Report. Available at: <http://www.publications.parliament.uk/pa/cm200405/cmselect/cmdfence/63/6306.htm> [Accessed 29 October 2015].

Defence Suppliers Directory, n.d. *Warrior Armoured Infatry Fighting Vehicle FV 510.* [online] armedforces.co.uk Defence Suppliers Directory. Available at: <http://www.armedforces.co.uk/armyindex.php> [Accessed 26 August 2015].

Drullman, R., 1995. Temporal envelope and fine structure cues for speech intelligibility. *The Journal of the Acoustical Society of America*, 97(1), p. 585–592.

EASHW (The European Agency for Safety and Health at Work), 2009. *Combined exposure to noise and ototoxic substances*. [pdf] European Risk Observatory. Available at: <https://osha.europa.eu/en/publications/literature_reviews/combined-exposure-to-noise-and-ototoxic-substances> [Accessed 08 September 2015].

Eddins, D. A. & Liu, C., 2012. Psychometric properties of the coordinate response measure corpus with various types of background interference. *Journal of the Acoustical Society of America,* 31(2), p.177–83.

Encyclopaedia Britannica, 2015. *Encyclopaedia Britannica Dialect Article*. [online] London: Encyclopedia Britannica (UK). Available at: <http://www.britannica.com/EBchecked/topic/161156/dialect> [Accessed 26 August 2015].

Endsley, M., 1995. Towards a theory of Situational Awareness in Dynamic Systems. *Human Factors*, 37(1), p.32–64.

Equality Act, 2010. *Equality Act 2010 Chapter 15*. [online] London: Stationary Office. Available at: <http://www.legislation.gov.uk/ukpga/2010/15/contents> [Accessed 26 August 2015].

References

Erlandsson, B., Hàkanson, H., Ivarsson, A., Nilsson, P. & Wersäll, J., 1980. Hair cell damage in the guinea pig due to different kinds of noise. *Acta Oto-laryngologica Supplement*, 367, p.1–43.

Espaillat, F. A. & Smith, J.A., 2010. Combined arms tactical trainers: Supporting the reserve component. [online] National Guard Article View. Available at: <http://www.nationalguard.mil/News/ArticleView/tabid/5563/Article/586718/combined-arms-tactical-trainers-supporting-the-reserve-component.aspx> [Accessed 27 August 2015].

European Commission, 2008. *Potential health risks of exposure to noise from personal music players and mobile phones including a music playing function.* [pdf] Scientific Committee on Emerging and Newly Identified Health Risks. Available at: <http://ec.europa.eu/health/ph_risk/committees/04_scenihr/docs/scenihr_o_017.pdf> [Accessed 26 August 2015].

European Parliament, 2003. *Directive 2003/10/EC on the minimum health and safety requirements regarding the exposure of workers to the risks arising from physical agents*.[online] European Agency for Safety and Health at Work. Available at: <https://osha.europa.eu/en/legislation/directives/82> [Accessed 26 August].

Faul, F., Erdfelder, E., Lang, A.G. & Buchner, A., 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behaviour Research Methods*, 39(2), p.175–91.

Field, A., 2005a. *Discovering Statistics Using SPSS.* 2nd ed. London: Sage Publications.

Field, A., 2005b. *Effect Sizes*. [pdf] Statistics Hell. Available at: <http://www.statisticshell.com/docs/effectsizes.pdf> [Accessed 26 August 2015].

Fleiss, J., Levin, B. & Paik, M., 2003. *Statistical methods for rates and proportions*. 3rd ed. Hoboken: John Wiley and Sons.

Fletcher, H., 1940. Auditory Patterns. *Reviews of Modern Physics*, 12, p.47–65.

Fletcher, H., 1950. A method for calculating hearing for speech from the audiogram. *The Journal of the Acoustical Society of America*, 22(1), p. 1-5.

Forshaw, S.E., Ritmiller, L.M., Laroche, C., Hodgson, M., Hamilton, K. & Cameron, B.J., 1999. *Hearing Standards for Seagoing Personnel* [pdf]. Ergonomics and Human Factors Groups. Available at: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.127.1395&rep=rep1&type=pdf> [Accessed 26 August 2015].

Foster, J.R. & Haggard, M.P., 1987. The four alternative auditory feature test (FAAF)--linguistic and psychometric properties of the material with normative data in noise. *British Journal of Audiology,* 21(3), p.165–74.

Fulcher, E., 2003. *Cognitive Psychology Crucial Study Guide Series*. Exeter: Learning Matters.

Gatehouse, S., Naylor, G. & Elberling, C. 2003. Benefits from hearing aids in relations to the interaction between the user and the environment. *International Journal of Audiology*, 42 (Suppl. 1), p. S77-85.

Giguère, C., Laroche, C., Soli, S.D. & Vaillancourt, V., 2008. Functionally-based screening criteria for hearing-critical jobs based on the Hearing in Noise Test. *International Journal of Audiology,* 47(6), p.319–28.

Gnansia, D., Péan, V., Meyer, B., & Lorenzi, C., 2009. Effects of spectral smearing and temporal fine structure degradation on speech masking release. *The Journal of the Acoustical Society of America*, 125(6), p.4023–4033.

Gordan-Salant, S. & Fitzgibbons, P.J., 1997. Selected Cognitive Factors and Speech Recognition Performance Among Young and Elderly Listeners. *The Journal of Speech, Language and Hearing Research,* 40, p.423-31.

Grantham, M.A.M., 2012. Noise- Induced Hearing Loss and Tinnitus Challenges for the Military. In C.G. Le Prell, D. Henderson, R.R. Fay & A.N. Popper, eds. 2012. *Noise- Induced Hearing Loss Scientific Advances*. New York: Springer. Chapter 3.

Hall, S.J., 2006. *The development of a new English sentence in noise test and an English number recognition test.* Ph. D. University of Southampton.

Harris, J.D., 1965. Pure-Tone Acuity and the Intelligibility of Everyday Speech. *Journal of the Acoustical Society of America,* 37(5), p.824–30.

Health and Safety Executive, 2014. *Frequently asked questions: What are risk matrices?* [online] Health and Safety Executive Guidance. Available at: <http://www.hse.gov.uk/risk/faq.htm> [Accessed 27 August 2015].

HearCom, 2006. *FP6- 004171 HEARCOM Hearing in the Communication Society D-1-3 : Protocol for implementation of communication tests in different languages.* [pdf] HearCom. Available at: <http://hearcom.eu/about/DisseminationandExploitation/deliverables/HearCom_D1-3_V2-0.pdf> [Accessed 27 August 2015].

Heinrich, A., Henshaw, H. & Ferguson, M.A., 2015. The relationship of speech intelligibility with hearing sensitivity, cognition, and perceived hearing difficulties varies for different speech perception tests. *Frontiers in Psychology,* 6(782), p. 1-14.

Hemingway, P. & Brereton, N., 2009. *What is a systematic review?* [online] What is…? Series. Available at: <http://www.medicine.ox.ac.uk/bandolier/painres/download/whatis/syst-review.pdf> [Accessed 13 November 2015].

Henderson, D. & Hamernik, R., 1995. Biologic basis of noise-induced hearing loss. *Occupational Medicine,* 10(3), p.513-34.

Henderson, D. & Hamernik, R.P., 1986. Impulse noise: critical review. *Journal of the Acoustical Society of America,* 80(2), p.569–84.

Hopkins, K., & Moore, B. C. J., 2009. The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise. *The Journal of the Acoustical Society of America*, 125(1), p.442–446.

Hopkins, K., Moore, B. C., & Stone, M. C., 2008. Effects of moderate cochlear hearing loss on the ability to benefit from temporal fine structure information in speech. *The Journal of the Acoustical Society of America*, 123(2), p.1140–1153.

Hopkins, W. G., 2000. Measures of reliability in Sports Medicine and Science. *Sports Medicine,* 30(1), p.1-15.

Hu, B., 2012. Noise- Induced Structural Damage to the Cochlea. In C.G. Le Prell, D. Henderson, R.R. Fay & A.N. Popper, eds. 2012. *Noise- Induced Hearing Loss Scientific Advances*. New York: Springer. Chapter 5.

References

Hu, B.H., Guo, W., Wang, P.Y., Henderson, D., Jiang, S.C., 2000. Intense noise-induced apoptosis in hair cells of guinea pig cochleae. *Acta Oto-laryngologica*, 120(1), p.19–24.

Human Experimentation Safety and Ethics Committee, 1996. *Guide to Experimentation involving Human Subjects ISVR Technical Memorandum No 808.* Southampton: Institute of Sound and Vibration Research.

Humes, L., Joellenbeck, M.L. & Durch, J.S., 2006. Noise and Noise-Induced Hearing Loss in the Military. In L. Humes, M.L. Joellenbeck, & J.S. Durch, eds. 2006. *Noise and Military Service.* Washington: National Academies Press. Chapter 3.

Humes, L.E. & Roberts, L., 1990. Speech-recognition difficulties of the hearing-impaired elderly: the contributions of audibility. *Journal of Speech and Hearing Research*, 33(4), p.726–35.

Humes, L.E. 2002. Factors underlying the speech-recognition performance of elderly hearing-aid wearers. *Journal of the Acoustic Society of America,* 112, p.1112-32.

Humes, L.E., Burk, M.H., Coughlin, M.P., Busey, T.A., & Strauser, L.E. 2007. Auditory speech recognition and visual text recognition in younger and older adults: Similarities and differences between modalities and the effects of presentation rate. *Journal of Speech, Language, and Hearing Research,* 50(2), p.283-303.

Jamieson, S., 2004. Likert scales: how to (ab)use them. *Medical Education*, 38, p.1212–18.

Johansson, M.S. & Arlinger, S.D., 2002. Binaural masking level difference for speech signals in noise. *International Journal of Audiology*, 41(5), p.279–84.

Joris, P.X., 2009. Recruitment of Neurons and Loudness. *Journal of the Association for Research in Otolaryngology*, 10(1) p.1-4.

Khan Academy, 2014. *Calculus: Derivatives- Introduction to derivatives*. [online] Available at: <https://www.khanacademy.org/math/differential-calculus/taking-derivatives/derivative_intro/v/calculus-derivatives-1> [Accessed 01 September 2015].

Khan Academy, 2015. *R-sqaured or coefficient of determination*. [online] Available at: <https://www.khanacademy.org/math/probability/regression/regression-correlation/v/r-squared-or-coefficient-of-determination> [Accessed June 9, 2015].

Kidd, G.R., Watson, C.S. & Gygi, B. 2007. Individual differences in auditory abilities. *Journal of the Acoustical Society of America,* 122(1), p.418-35.

King, A.V.M., Coles, R.R.A., Lutman, M.E. & Robinson, D.W., 1992. *Guidelines for Medicolegal Practice: Assessment of Hearing Disability*. London: Whurr Publishers.

Kingdom, F. & Prins, N., 2010. *Psychophysics: A Practical Introduction.* London: Elsevier Limited.

Kinson, T., 2012. *The tech behind the check*. [online] Available at: <http://www.actiononhearingloss.org.uk/community/blogs/our-guest-blog/the-tech-behind-the-check.aspx> [Accessed 01 September 2015].

Kitterick, P.T., Bailey, P.J. & Summerfield, A.Q., 2010. Benefits of knowing who, where, and when in multi-talker listening. *Journal of the Acoustical Society of America*, 127(4), p.2498–508.

Kitzinger, J., 1995. Qualitative Research: Introducing focus groups. *British Medical Journal*, 311(7000), p.299–302.

Kocsis, J.D. & Tessler, A., 2009. Pathology of blast-related brain injury. Journal of Rehabilitation Research and Development, 46(6), p.667–72.

Kollmeier, B. & Wesselkamp, M., 1997. Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment. *Journal of the Acoustical Society of America*, 102(4), p.2412–21.

Kramer, S.E., Kapteyn, T.S., Festen, J.M. & Tobi, H., 1996. The Relationships between Self-reported Hearing Disability and Measures of Auditory Handicap. *International Journal of Audiology*, 35(5), p.277–87.

Kujawa, S.G. & Liberman, M.C., 2009. Adding Insult to Injury: Cochlear Nerve Degeneration after "Temporary" Noise-Induced Hearing Loss. *Journal of Neuroscience,* 29(45), p.14077–85.

Laerd Statistics, n.d. *One-way ANOVA in SPSS Statistics.* [online] Available at: <https://statistics.laerd.com/spss-tutorials/one-way-anova-using-spss-statistics-2.php> [Accessed 26 October 2015].

Laroche, C., Giguère, C., Soli, S.D. & Vaillancourt, V., 2008. Establishment of fitness standards for hearing-critical jobs. In: ICBEN (International Commission on Biological Effects of Noise), *9th International Congress on Noise as a Public Health Problem*. Mashantucket, United States. 21-25 July 2008.

Laroche, C., Soli, S., Giguere, C., Vaillancourt, V. & Fortin, M., 2003. An approach to the development of hearing standards for hearing-critical jobs. *Noise and Health*, 6(21), p.17–37.

Lee, I.A. & Preacher, K.J., 2013. *Calculation for the test of the difference between two dependent correlations with one variable in common*. [online computer software] Available at: <http://quantpsy.org/corrtest/corrtest2.htm> [Accessed 02 September 2015].

Leek, M.R., 2001. Adaptive procedures in psychophysical research. *Perception and Psychophysics*, 63(8), p.1279–92.

Leensen, M.C., 2013. *Noise induced hearing loss: Screening with pure-tone audiometry and speech-in-noise testing*. Ph. D. University of Amsterdam.

Leensen, M.C., de Laat, J. A. & Dreschler, W.A., 2011a. Speech-in-noise screening tests by internet, part 1: test evaluation for noise-induced hearing loss identification. *International Journal of Audiology*, 50(11), p.823–34.

Leensen, M.C., de Laat, J. A., & Snik, A.F. & Dreschler, W.A., 2011b. Speech-in-noise screening tests by internet, Part 2: Improving test sensitivity for noise-induced hearing loss. *International Journal of Audiology*, 50(11), p.835–48.

Léger, A. ., Moore, B. C. J., & Lorenzi, C., 2012. Abnormal speech processing in frequency regions where absolute thresholds are normal for listeners with high-frequency hearing loss. *Hearing Research*, 294(1-2), p.95-103.

Levitt, H., 1971. Transformed Up-Down Methods in Psychoacoustics. *Journal of the Acoustical Society of America*, 49(2), p.467-77.

Lockheed Martin, 2015. *UK Combined Arms Tactical Trainer (UK CATT)*. [online] Available at: <http://www.lockheedmartin.co.uk/uk/what-we-do/products/UKCombinedArmsTacticalTrainer.html> [Accessed 02 September 2015].

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S. & Moore, B.C., 2006. Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. *Proceedings of the National Academy of Sciences of the United States of America,* 103(49), p.18866–9.

References

Lower, M., 2014. *Technical information about the University of Southampton small anechoic chamber.* [email] (Personal Communication, 19 November 2014).

Lunner, T. & Sundewall-Thorén, E. 2007. Interactions between cognition, compression, and listening conditions: Effects on speech-in-noise performance in a two-channel hearing aid. *Journal of the American Academy of Audiology*, 18(7), p.604-17.

Lutman, M.E., Hall, S.J. & Athalye, S., 2006. Development of a telephone hearing test. *Proceedings of the Institute of Acoustics,* 28(1), p.240–43.

Ma, J., Hu, Y. & Loizou, P.C., 2009. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *Journal of the Acoustical Society of America,* 125(5), p.3387–405.

MATLAB (Matrix Laboratory), 2010. MATLAB 7.10.0. [computer program] The MathWorks Incorporation, Natick, Massachusetts. Available at: <http://uk.mathworks.com/products/matlab/index.html> [Accessed 02 September 2015].

Mattys, S.L., Davis, M.H., Bradlow, A.R. & Scott, S.K. 2012. Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, 27(7/8), p.953-78.

Mayo, L. H., Florentine, M., & Buss, S. 1997. Age of second-language acquisition and perception of speech in noise. *Journal of Speech, Language, and Hearing Research*, 40, p.686-93.

Mayorga, M.A., 1997. The pathology of primary blast overpressure injury. *Toxicology*, 121(1), p.17–28.

McLeod, S.A., 2007. *What is Validity.* [online] Simply Psychology. Available at: <http://www.simplypsychology.org/validity.html> [Accessed 03 September 2015].

Meggitt, 2015. *Military training solutions*. [online]. Meggitt Training Systems. Available at: <http://www.meggitttrainingsystems.com/military.aspx > [Accessed 03 September 2015].

Meyer Sound, n.d. *Speech Intelligibility Papers Glossary of Terms- Modified Rhyme Test*. [online] Available at: <http://www.meyersound.com/support/papers/speech/glossary.htm#mrt> [Accessed 01 September 2015].

Ministry of Defence (2014). *JSP (Joint Services Publication) 950: Medical policy: Part 1, Vol 6 Ch 4. Leaflet 6-4-2: Assessing Audiograms – Guidance for Medical Staff*.

Moon, I. J., & Hong, S. H., 2014. What Is Temporal Fine Structure and Why Is It Important? *Korean Journal of Audiology*, 18(1), p.1–7.

Moore B.C.J. & Glasberg B.R., 1997. A model of loudness perception applied to cochlear hearing loss. *Auditory Neuroscience*, 3, p.289–311.

Moore, B. C. J., 2002. Psychoacoustics of normal and impaired hearing. *British Medical Bulletin*, 63(1), p.121–134.

Moore, B.C.J. & Glasberg, B.R., 1993. Simulation of the effects of loudness recruitment and threshold elevation on the intelligibility of speech in quiet and in a background of speech. *Journal of the Acoustical Society of America,* 94(4), p.2050–62.

Moore, B.C.J., 1998. Speech Perception. In B.C.J. Moore, 1996. *Cochlear Hearing Loss*. 2nd ed. London: Whurr Publishers. Chapter 8.

Moore, B.C.J., 2003. Speech processing for the hearing-impaired: successes, failures, and implications for speech mechanisms. *Speech Communication,* 41(1), p.81–91.

Moore, B.C.J., 2008a. *An Introduction to the Psychology of Hearing*. 5th ed. Bingley: Emerald Group Publishing Limited.

Moore, B.C.J., 2008b. The Role of Temporal Fine Structure Processing in Pitch Perception, Masking, and Speech Perception for Normal-Hearing and Hearing-Impaired People. *Journal of the Association for Research in Otolaryngology*, 9(4), p.399–406.

Muhr, P., 2010. *Hearing in young men- The influence of military noise and epidemiological aspects*. Ph. D. Karolinska Institutet.

Na´belek, A. K., & Donahue, A. M. 1984. Perception of consonants in reverberation by native and non-native listeners. *Journal of the Acoustical Society of America*, 75, p.632-34.

Nath, A. R., & Beauchamp, M. S., 2012. A Neural Basis for Individual Differences in the McGurk Effect, a Multisensory Speech Illusion. *Neuroimage*, 59(1), p.781–787.

NATO (North Atlantic Treaty Organisation), 2010. *Technical Report on Hearing Protection- Needs, Technologies and Performance* (TR-HRM-147). [pdf] Research and Technology Organisation Available at: <http://www.dtic.mil/dtic/tr/fulltext/u2/a539790.pdf> [Accessed 03 September 2015].

Nilsson, M., Soli, S.D. & Sullivan, J.A., 1994. Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. *Journal of the Acoustical Society of America,* 95(2), p.1081–99.

NIOSH (National Institute for Occupational Safety and Health), 2001. *Work Related Hearing Loss* (2001-103). [pdf] Centre for Disease Control and Prevention. Available at: <http://www.cdc.gov/niosh/docs/2001-103/#1> [Accessed October 8, 2012].

Nottingham Hearing Biomedical Research Unit. 2015. *A James Lind Alliance Priority Setting Partnership for Mild-Moderate Hearing Loss.* [online] NHS Institute for Health Research. Availble at: <http://www.hearing.nihr.ac.uk/research/a-james-lind-alliance-priority-setting-partnership-for-mild-moderate-hearin/> [Accessed 17 November 2015].

Oxenham, A. J., & Bacon, S. P., 2003. Cochlear Compression: Perceptual Measures and Implications for Normal and Impaired Hearing. *Ear and Hearing*, 24(5), p.352–366.

Oxford English Dictionary, 2015. *Oxford English Dictionary Online.* [online] Oxford University Press (UK). Available through: <www.oed.com> [Accessed 03 September 2015].

Ozimek, E., Kutzner, D., Sęk, A. & Wicher, A., 2009. Development and Evaluation of the Polish Triplet Test. *Speech Communication*, 51(4), p.307–16.

Payne, W. & Harvey, J., 2010. A framework for the design and development of physical employment tests and standards. *Ergonomics*, 53(7), p.858–71.

Pearn, M. & Kandola, R., 1988. *Job Analysis: A Practical Guide for Managers*. London: Institute of Personnel Management.

Pearson, C., 2011. The Characteristics of Pure-tone Audiograms in a Sample of Royal Marines After Operation Herrick 9. *Journal of the Royal Naval Medical Service,* 97(3), p.123–6.

Pedersen, E.R. & Juhl, P.M., 2014. User-operated speech in noise test: Implementation and comparison with a traditional test. *International Journal of Audiology,* 53(5), p.336–44.

References

Persson, P., Harder, H., Arlinger, S. & Magnuson, B., 2001. Speech recognition in background noise: monaural versus binaural listening conditions in normal-hearing patients. *Otology and Neurotology,* 22(5), p.625–30.

Peters, R.W., Moore, B.C.J. & Baer, T., 1998. Speech reception threshold in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people. *Journal of the Acoustical Society of America,* 103(1), p.577–87.

Plack, C.J., Barker, D. & Prendergast, G., 2014. Perceptual consequences of "hidden" hearing loss. *Trends in Hearing,* 18, p.1–11.

Plomp, R., 1978. Auditory handicap of hearing impairment and the limited benefit of hearing aids. . *Journal of the Acoustical Society of America,*63(2), p.533-49.

Plomp, R., 1986. A signal-to-noise ratio model for the speech reception threshold of the hearing impaired. *Journal of Speech and Hearing Researc*h, 29(2), p.146–54.

Prins, N. & Kingdom, F., 2009. *Palamedes: Matlab routines for analyzing psychophysical data.* [computer software] Available at: <http://www.palamedestoolbox.org> [Accessed 03 September 2015].

QSR International Pty Ltd, 2012. *NVivo qualitative data analysis software (Version 10)*. [computer software] Available at: <http://www.qsrinternational.com/default.aspx> [Accessed 03 September 2015].

Quintana, A. & Pérez, G., 2003. Stopping rules in adaptive Bayesian threshold estimation. In IX Conferencia Española de Biometría. In Conference Proceedings for the Spanish Conference on Biometrics, *9th Spanish Conference on Biometrics.* A Coruña, Spain. 28-30 May 2003.

RandomOrg, 2014. List Randomiser. [online computer software] Available at: <http://www.random.org/> [Accessed 04 September 2015].

Reed, C.M., Braida, L.D. & Zurek, P.M., 2009. Review of the Literature on Temporal Resolution in Listeners with Cochlear Hearing Impairment: A Critical Assessment of the Role of Suprathreshold Deficits. *Trends in Amplification*, 13(1), p.4–43.

Rhebergen, K.S. & Versfeld, N.J., 2005. A speech intelligibility index-based approach to predict the speech reception thresholds for sentences in fluctuating noise for normal-hearing listeners. *Journal of Acoustical Society of America,* 117(4), p.2181–92.

Rogers, C. L., Lister, J. J., Febor, D. M., Besing, J. M., & Abrams, H. B. 2006. Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing. *Applied Psycholinguistics*, 27(3), p.465-85.

Rosen, S., Souza, P., Ekelund, C. & Majeed, A.A., 2013. Listening to speech in a background of other talkers : Effects of talker number and noise vocoding. *Journal of Acoustical Society of America,* 133(4), p.2431–43.

Rothpletz, A.M., Wightman, F.L. & Kistler, D.J., 2011. Informational Masking and Spatial Hearing in Listeners With and Without Unilateral Hearing Loss. *Journal of Speech Language and Hearing Research*, 55(2), p.511–31.

Rudner, M., Foo, C., Sundewall-Thorén, E., Lunner, T. & Rönnberg, J. 2008. Phonological mismatch and explicit cognitive processing in a sample of 102 hearing-aid users. *International Journal of Audiology*, 47 (Suppl. 2), p.S163-70.

Scharine, A.A., Letowski, T.R. & Sampson, J.B., 2009. Auditory Situation Awareness in Urban Operations. *Journal of Military Strategic Studies,* 11(4), p.1–24.

Schoepflin, J.R., 2012. *Back to Basics: Speech Audiometry*. [online] Audiology Online. Available at: <http://www.audiologyonline.com/articles/back-to-basics-speech-audiometry-6828> [Accessed 08 September 2015].

Semeraro, H.D., Bevis, Z.L., Rowan, D., van Besouw, R.M &, Allsopp, A.J., 2015. Fit for the frontline? Identification of mission-critical auditory tasks (MCATs) carried out by infantry and combat-support personnel. *Noise and Health*, 17(75), p.98–107.

Shuttleworth, M., 2009a. *External validity*. [online] Explorable. Available at: <https://explorable.com/external-validity> [Accessed 08 September 2015].

Shuttleworth, M., 2009b. *Population validity*. [online] Explorable. Available at: <https://explorable.com/population-validity> [Accessed 08 September 2015].

Shuttleworth, M., 2009c. *Types of validity*. [online] Explorable. Available at: <https://explorable.com/statistical-validity> [Accessed 08 September 2015].

Smits, C., Kapteyn, T.S. & Houtgast, T., 2004. Development and validation of an automatic speech-in-noise screening test by telephone. *International Journal of Audiology*, 43(1), p.15–28.

Smoorenburg, G.F., 1992. Speech reception in quiet and in noisy conditions by individuals with noise-induced hearing loss in relation to their tone audiogram. *Journal of the Acoustical Society of America*, 91(1), p.421–37.

Spoendlin, H., 1971. Primary structures changes in the Organ of Corti after acoustic overstimulation. *Acta Oto-Laryngologica*, 71(1-6), p.166–76.

Stanton, N.A., Chambers, P.R.G. & Piggott, J., 2001. Situational Awareness and Safety. *Safety Science*, 39(3), p.189–204.

Steiger, J.H., 1980. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87(2), p.245–51.

Strayer, D.L. & Johnston, W.A., 2001. Driven to Distraction: Dual-Task Studies of Simulated Driving and Conversing on a Cellular Telephone. *Psychological Science*, 12(6), p.462-66.

Summerfield, Q., Palmer, A.R., Foster, J.R., Marshall, D.H. & Twomey, T., 1994. Clinical evaluation and test-retest reliability of the IHR-McCormick Automated Toy Discrimination Test. *British Journal of Audiology*, 28(3), p.165–79.

Surprenant, A.M. & Watson, C.S. 2001. Individual differences in the processing of speech and nonspeech sounds by normal-hearing listeners. *Journal of the Acoustical Society of America,* 110(4), p.2085-95.

Taylor, B., 2003. Speech-in-noise tests: How and why to include them in your basic test battery. *The Hearing Journal,* 56(1), p.40–44.

The Department for Health, 2015. *Action Plan on Hearing Loss.* [pdf] NHS England and Department of Health. Available at: < https://www.england.nhs.uk/wp-content/uploads/2015/03/act-plan-hearing-loss-upd.pdf> [Accessed 13 November 2015].

The Royal British Legion, 2014. *Lost Voices*. [pdf] The Royal British Legion. Available at: <http://www.britishlegion.org.uk/media/2282/lostvoiceshearinglossreport.pdf> [Accessed 08 September 2015].

References

The Telegraph, 2015. *Police marksman removed from duties after failing hearing test wins discrimination claim.* [online] The Telegraph Newspaper. Available at: <http://www.telegraph.co.uk/news/uknews/law-and-order/11903697/Police-marksman-removed-from-duties-after-failing-hearing-test-wins-discrimination-claim.html> [Accessed 30 November 2015].

Tufts, J.B., 2011. *Fitness for Duty*. [online] Council for Accreditation in Occupational Hearing Conservation. Available at: <http://www.caohc.org/updatearticles/spring2011/fitness_for_duty.php?mode=print> [Accessed 08 September 2015].

Tufts, J.B., Vasil, K. A. & Briggs, S., 2009. Auditory fitness for duty: a review. *Journal of the American Academy of Audiology*, 20(9), p.539–57.

University of Southampton, 2013. *Risk Assessment Template*. [online] Safety and Occupational Health. Available at: <http://www.southampton.ac.uk/healthandsafety/risk_assessment/> [Accessed 08 September 2015].

Van Rooij, J.C.G.M., Plomp, R. & Orlebeke, J.F. 1989. Auditive and cognitive factors in speech perception by elderly listeners. I: Development of test battery. *Journal of the Acoustical Society of America,* 86, p.1294-309.

Versfeld, N.J. & Dreschler, W.A., 2002. The relationship between the intelligibility of time compressed speech and speech in noise in young and elderly listeners. *Journal of the Acoustical Society of America*, 111(1 Pt 1), p.401–8.

Wagener, K., Josvassen, J.L. & Ardenkjaer, R., 2005. Design, optimization, and evaluation of a Danish sentence test in noise. *International Journal of Audiology*, 42(1), p.10–17.

Wallace, D., 2012. *Black Jack Brigade field tests new training systems*. [online] Fort Hood Sentinel. Available at: <http://www.forthoodsentinel.com/story.php?id=10003> [Accessed 08 September 2015].

Weisstein, E., 2014. *Least Squares Fitting*. [online] MathWorld. Available at: <http://mathworld.wolfram.com/LeastSquaresFitting.html> [Accessed 08 September 2015].

WHO (World Health Organisation), 2002. *Chapter 4: Quantifying Selected Major Risks to Health*. [pdf] The World Health Report Reducing Risks, Promoting Healthy Life. Available at: <http://www.who.int/whr/2002/en/whr02_en.pdf> [Accessed 08 September 2015].

Wilson, R.H. & Cates, W.B., 2008. A comparison of two word-recognition tasks in multitalker babble: speech Recognition in Noise Test (SPRINT) and Words-in-Noise Test (WIN). *Journal of the American Academy of Audiology,* 19(7), p.548–56.

Wilson, R.H., McArdle, R.A. & Smith, S.L., 2007. An Evaluation of the BKB-SIN, HINT, QuickSIN, and WIN Materials on Listeners With Normal Hearing and Listeners With Hearing Loss. *Journal of Speech, Language and Hearing Research*, 50(4), p.844–56.

Xu, L., Thompson, C. S., & Pfingst, B. E., 2005. Relative contributions of spectral and temporal cues for phoneme recognition. *The Journal of the Acoustical Society of America*, 117(5), p.3255–3267.

Yang, W.P., Henderson, D., Hu, B.H. & Nicotera, T.M., 2004. Quantitative analysis of apoptotic and necrotic outer hair cells after exposure to different levels of continuous noise. *Hearing Research*, 196(1-2), p.69–76.

Ylikoski, M.E. & Ylikoski, J.S., 1994. Hearing loss and handicap of professional soldiers exposed to gunfire noise. *Scandinavian Journal of Work, Environment and Health,* 20(2), p.93–100.

Zekveld, A. A., Kramer, S. E., & Festen, J. M., 2011. Cognitive Load During Speech Perception in Noise: The Influence of Age, Hearing Loss, and Cognition on the Pupil Response. *Ear and Hearing*, 32(4), p.498–510.

Zekveld, A.A., Heslenfeld, D.J., Festen, J.M. & Schoonhoven, R. 2006. Top-down and bottom-up processes in speech comprehension. *NeuroImage,* 32, p.1826-36.

Zhao, F. & Stephens, D., 2007. A critical review of King-Kopetzky syndrome: Hearing difficulties, but normal hearing? *Audiological Medicine*, 5(2), p.119-24.

Żychaluk, K. & Foster, D.H., 2009. Model-free estimation of the psychometric function. *Attention, perception and psychophysics,* 71(6), p.1414–25.