# University of Southampton

FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

Institute of Sound and Vibration Research

Developing an Arabic speech in noise test as a measure of auditory fitness for duty

by

**IMAN OSAMAH RAWAS**

ORCID ID [0000-0003-4846-8908]

Thesis for the degree of Doctor of Philosophy

April 2023

# University of Southampton

## Abstract

FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

Institute of Sound and Vibration research

Thesis for the degree of Doctor of Philosophy

Developing an Arabic speech in noise test as a measure of auditory fitness for duty

by

Iman Osamah Rawas

Hearing commands in noisy backgrounds is a frequent and important task in many fast-response occupations such as the military and the police force. Inability to hear and adequately respond may have detrimental effects on employees and their co-workers. Measuring auditory fitness for duty (AFFD) accurately is critical for employing organizations and the livelihood of their employees. While pure-tone audiometry (PTA) is the mainstay of AFFD assessment in Saudi Arabia, it is not a valid predictor of the ability to hear in noise. Speech-in-noise (SiN) tests have higher face validity as AFFD predictors for jobs relying on communication in noise. Arabic SIN tests are scarce, and none have been developed or used for AFFD purposes. Many factors affect SIN test performance. Personality trait conscientiousness, which has been shown to be positively related to job task-performance, may be an overlooked factor related to auditory task performance. The aims of this research were to 1) develop and equalize speech material in noise suitable for fast-response occupations, ensuring homogeneity of the speech material and to adapt the developed material into a test suitable for a representative occupation; and 2) explore the discriminative ability of the developed test to detect mild sensorineural hearing impairment and factors affecting test performance under standard and more challenging listening conditions.

An Arabic speech corpus was developed and optimized for intelligibility in noise. The acceptable levels of variation in homogenous speech material under conditions similar to the developed test were explored using Monte Carlo simulations. The speech material was then implemented into an adaptive procedure and named the Arabic commands in noise test (ACINT). The resultant test was a general SiN test with military characteristics. To explore the effect of conscientiousness on the representative population chosen in the early stages of test development, the Royal Saudi Air Defence (RSADF), feasibility of the personality measure chosen for use, the Arabic NEO-FFI, was determined on Saudi military and civilian samples representative of the populations. Performance on the ACINT was then explored in a normal-hearing sample representative of entry-level military recruits and a hearing-impaired sample, in different listening conditions. There was no effect of conscientiousness in performance as assessed in normal-hearing individuals. The ACINT in its standard format was found to have good repeatability. It also showed good sensitivity and specificity in discriminating between normal and hearing-impaired

individuals. Future work is required to assess test reliability and assign cut-off points to the test, based on job specific criterion obtained from large scale task- and noise environment-analysis studies of the target population.

# Table of Contents

# Table of Tables

# Table of Figures

# Research Thesis: Declaration of Authorship

Print name: IMAN OSAMAH RAWAS

Title of thesis: Developing an Arabic speech in noise test as a measure of auditory fitness for duty

I 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.

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. None of this work has been published before submission;

Signature: ........................................................................................ Date: ...........................

# Acknowledgements

I would like to thank all the cooperating institutions, the Saudi Ministry of Education, and the Royal Saudi Air Defence Academy for facilitating this research. I would like to thank King Abdul-Aziz University for sponsoring me, with special thanks to the joint supervision programme. Ms. Umaima, your support was appreciated. Thank you to my colleague, Mohammed Numaan for helping me secure equipment for my experiments. Thank you to all the people who participated in the studies.

I would like to thank all those who supervised me at any point during this journey. Dr Daniel Rowan, thank you for teaching me critical thinking and saving me with code. Dr Hannah Semeraro, thank you for teaching me to be less hasty. Dr Victoria Watson, thank you for your forever positive and constructive feedback. Dr Stefan Bleeck, thank you for boosting my confidence and guiding me to be more meticulous in the end. A special thanks to my internal supervisor, Dr Afaf Bamanie, for her kind motherly guidance and support throughout this journey.

Where do I even begin with my family. Thank you to my brothers and sister for each contributing in their own special way and my mother for providing constant childcare and prayers. Mama, I couldn’t have done this without you. Thank you, Judy, Aminah and Hamza, for helping me win the PhD challenge, as they like to call it. I am forever grateful to my biggest supporter, my husband Hattan, for believing in me even when I did not believe in myself and cheering me on.

Last but not least, I would like to thank my sister, confidante and PhD partner Sarah. We have been through thick and thin; I don’t know how I could have done it without you.

I dedicate this work to my father, may he rest in peace, he always set the achievement bar high.

# Abbreviations

| | |
|---|---|
| **ACINT** | Arabic commands in noise test |
| **AFFD** | Auditory fitness for duty |
| **CRM** | Coordinate response measure |
| **CVC** | Consonant-vowel-consonant |
| **dB A** | Decibel A-weighted curve |
| **DFO** | Department of Fisheries and Oceans |
| **DTT** | Digit Triplet Test |
| **EHF** | Extended high frequency |
| **FFD** | Fitness for duty |
| **FFM** | Five factor model |
| **FFT** | Five factor theory |
| **GPT** | Generic predictive test |
| **GUI** | Graphical user interface |
| **HCP** | Hearing conservation programme |
| **HCT** | Hearing critical task |
| **HINT** | Hearing in noise test |
| **HI** | Hearing impaired |
| **HL** | Hearing loss |
| **HPD** | hearing protection devices |
| **ICC** | Intra-class correlation coefficient |
| **IF** | Inferior colliculus |
| **ILD** | Inter-aural level difference |
| **ITD** | Inter-aural time difference |
| **LMM** | Linear mixed model |
| **LTASS** | Long term average speech spectrum |
| **MCS** | Monte Carlo simulation |
| **MILSINT** | Military sounds in noise |
| **MOC** | Medial olivocochlear |
| **MoCS** | Method of constant stimuli |
| **MRT** | Modified rhyme test |
| **MTBI** | Mild traumatic brain injury |
| **NEO-FFI** | NEO Five-Factor Inventory |
| **NEO-PI-R** | Revised NEO Personality Inventory |

| | |
|---|---|
| **NH** | Normal hearing |
| **NIHL** | Noise induced hearing loss |
| **NPV** | Negative predictive value |
| **NRR** | Noise reduction rating |
| **OEC** | Occupational Earcheck |
| **OHC** | Outer hair cell |
| **OHL** | Occupational hearing loss |
| **PES** | Physical employment standard |
| **PPV** | Positive predictive value |
| **PTA** | Pure tone audiometry |
| **RCMP** | Royal Canadian Mounted Police |
| **RCT** | Randomized controlled trial |
| **RMS** | Root mean square |
| **RSADF** | Royal Saudi Air Defence Forces |
| **SiN** | Speech in noise |
| **SME** | Subject matter expert |
| **SNHL** | Sensorineural hearing loss |
| **SNR** | Signal to noise ratio |
| **SOC** | Superior olivary complex |
| **SPRINT** | Speech Recognition in noise test |
| **SRT** | Speech recognition threshold |
| **SS** | Situational strength |
| **SSSN** | Stationary speech spectrum noise |
| **SVAN** | Salience/ventral attention network |
| **TFS** | Temporal fine structure |
| **TPT** | Task predictive tests |
| **TST** | Task simulation tests |
| **UK** | United Kingdom |
| **US** | United States |
| **WHO** | World Health Organization |
| **WM** | Working memory |

# Chapter 1 Introduction

## 1.1 Background and motivation

It is crucial to have good hearing in order to perform certain roles in some occupations, such as emergency services, law enforcement (Colaprete, 2012) and military services (St. Onge *et al.*, 2011; Semeraro *et al.*, 2015). In particular, roles commonly involving speech communication in background noise , like hearing commands in a background of running vehicles, machinery or multiple talkers (Laroche *et al.*, 2011; Bevis *et al.*, 2014; Soli, Amano-Kusumoto *et al.*, 2018). Consequently, hearing impairment can adversely affect an employee's job performance, depending on the role and the acoustical environment, potentially making the difference between life and death. An employer might therefore require the employee to pass a hearing assessment designed to check their hearing-related fitness for work, referred to as auditory fitness for duty (AFFD). Currently, pure-tone audiometry (PTA) is probably the most used AFFD test although there is doubt as to whether it can accurately predict AFFD for roles involving speech communication in noise. The inclusion of speech intelligibility in noise tests in AFFD testing has been recommended and has begun to be used in AFFD assessment standards (Vaillancourt *et al.*, 2011; Brammer and Laroche, 2012; Giguère *et al.*, 2019). Other recommendations are implementation of a complete test battery for auditory fitness functional hearing components, arguing the insufficiency of one test (Tufts *et al.*, 2018). Nevertheless, there is a consensus that PTA alone is not reflective of AFFD in most fast response jobs. Hence, the advocation for speech intelligibility in noise tests.

Currently, there is no speech in noise (SiN) test in Arabic for adults that has been validated and widely available. For most hearing-critical roles, there is also no reference auditory physical employment standard (PES) and consequently no direct way of determining whether a SiN test can distinguish between those with and without AFFD. It is therefore not clear whether a generic SiN test is suitable for all AFFD testing or whether a specific role requires a more tailored SiN test. The approach to designing AFFD assessments based on performance standards and development of suitable test measures is reliant on a rigorous analytical method.

The motivation behind this research was to develop and explore a measure for AFFD intended for fast-response jobs, focusing on a military sector. This PhD project initially set out to develop a general military-relevant SiN test and to incorporate it into an AFFD task that was to be designed based on a detailed task-analysis of the duties performed by the trainees of the Royal Saudi Air Defence Force (RSADF). This required frequent on-site presence on the RSADF premises and dynamic in-depth co-operation from the concerned target population. The COVID-19 pandemic prevented visits and interactions for almost a year. Following that, the RSADF were not willing to continue the previously agreed upon scope of co-operation and consented to a much more limited amount of participation. Considering these changes, a slightly different approach was taken towards test development. Based on general knowledge of military environments, and specific knowledge gleaned from our collaboration with the Royal Saudi Air Defence Force (RSADF), the test material, which had already been developed, was to be explored as a generic SiN test with potential for military populations that could also be used for more general audiological research. During the development process, factors affecting performance in SiN tests were further explored. A factor under-examined in the context of auditory task performance is personality trait conscientiousness. This trait is linked to improved task performance in academic and job settings (Judge *et al.*, 2008; Stacey and Kurunathan, 2015) but has yet to be explored in the context of AFFD.

The available literature on AFFD, current AFFD standards in different regions, and standards currently in use in Saudi Arabia were explored. The shortcomings of current measures were highlighted, and the first phase of the PhD was dedicated to developing an Arabic SiN measure to be studied as a general SiN with military relevant characteristics that could later be incorporated into an AFFD measure in certain occupations. For the second phase, the population previously chosen for exploration of the developed measure was further studied by personality assessments to explore aspects of test performance, focusing on the effect of personality construct conscientiousness and different population-relevant listening conditions. In the second study of phase two, a sample representative of the target population was assessed using the developed measure in different listening conditions to understand the ability of the measure to discriminate between normal hearing and hearing-impaired individuals, and the role of conscientiousness in test performance.

## 1.2 Research aims and objectives

The goal of this research was to develop and explore a measure for AFFD purposes in a Saudi population. This was achieved through two main aims across two phases.

**Aim 1** was to develop and validate a Saudi dialect-specific speech in noise test, the Arabic Commands in Noise Test (ACINT). **Phase I**, the developmental phase, addressed aim 1 through research **objectives one and two**.

**Research objective one**, addressed in chapter three, studies 1 & 2, was to develop and optimize generic predictive Arabic SiN test material potentially intended for occupations of a specific nature.

**Research objective two**, addressed in chapter four, study 3 was to explore the acceptable amount of variation in homogenous speech material using Monte Carlo Simulations.

**Aim 2** was to explore performance on the developed ACINT and assess its sensitivity to hearing loss. **Phase II**, the exploratory experimental phase, addressed aim 2 through **objectives three and four**.

**Research objective three**, chapter five, study 4, was to assess the feasibility of studying conscientiousness as a test performance predictor in a military population.

**Research objective four**, chapter six, study 5, was to assess the effect of different conditions and certain performer characteristics on test performance in generic and task-related conditions, by developing and exploring a task-related test for a particular role and assessing the discriminative ability of the generic ACINT compared to the ACINT in a task-related condition.

## 1.3 Thesis structure

This thesis is structured as follows:

**Chapter two** reviews the relevant literature. It provides an overview of hearing loss (HL) in the workplace and its effect on the employee. Auditory fitness for duty is reviewed. The process of hearing in noise, speech in noise tests used for occupational purposes, the role of SiN tests in AFFD assessments and the rationale for developing an Arabic SiN test are discussed. Factors affecting test performance are also reviewed and the rationale for choosing personality trait conscientiousness for further investigation as a factor affecting test performance is explained.

**Phase I** is covered in chapters three and four. **Chapter three** covers **(research objective one)**, developing Arabic speech material needed to develop a measure of AFFD ***(Study 1).*** Considerations for material development are discussed, and speech material development and equalization are explained. Further optimization of the test material utilizing an interleaving adaptive procedure ***(study 2***) is discussed. While developing the SiN test, a standard for the acceptable range of variation in homogenous speech material was not found. **Research objective two**, exploring 'how close is close enough' in terms of SiN test material homogeneity is explored using Monte Carlo simulations ***(study 3)*** and discussed in **chapter four**.

**Phase II** is covered in chapters five and six. the study was conducted with military populations in mind. Due to the nature of their job, they may differ from civilian populations in many aspects including personality expression. In **chapter five**, **research objective three**, the effect of conscientiousness in military and civilian populations and the feasibility of studying it as an affecting variable in a military population ***(study 4)*** are explored***.*** The feasibility of choosing a representative population based on differences and similarities in civilian and military personalities is discussed. **Chapter six** integrates the results of the previous studies through **research objective four**, exploring the performance in a normal-hearing representative population and a hearing-impaired population on developed speech material implemented into a task potentially representative of military communication conditions ***(study 5).*** The predictive ability of the test is assessed in different listening conditions relevant to the target population. The effect of conscientiousness on test performance in the normal-hearing population is also investigated.

## 1.4 Scientific contributions

The novel scientific contributions from this PhD are:

1. The design, development, and optimization of generic Arabic SiN test material that can be used for military audiology research and audiology research in other fast-response occupations.
2. The exploration of factors affecting Arabic speech material homogeneity and assessment of the amount of acceptable variation in homogenous speech material.
3. Insight into the ability of the developed SiN test to determine auditory performance on an auditory task in generic conditions compared to more challenging task-related conditions.
4. Knowledge regarding differences in expression of conscientiousness among Saudi military and civilian populations, and the effect of conscientiousness on performance in an auditory task.

# Chapter 2 Auditory fitness for duty

The goal of this research was to develop and explore a measure of AFFD. In this chapter, the concept of AFFD and its importance are explained, focusing on its place among occupational hearing assessments, the work that has been done in different regions by different researchers, current methods of assessment, their shortcomings and how this has been addressed. Following that, current AFFD practices specific to Saudi Arabia and gaps in knowledge are highlighted.

## 2.1 Hearing loss in the workplace

Communication is essential in most workplaces. To facilitate communication, individuals need a set of functional hearing abilities that vary depending on the nature of the job. Functional hearing abilities are those required for speech communication, maintaining awareness and functioning with the acoustic environment (Tufts, Vasil and Briggs, 2009; Al-Omari *et al.*, 2018). They are the ability to detect, recognize and localize sounds and the ability to recognize and understand speech. Speech comprehension is the most complex in the auditory skill hierarchy. It is categorized as an entity separate from other sounds because of its fundamentality for human communication and uniqueness to other signals (Tufts, Vasil and Briggs, 2009). Some have identified it as the most important hearing ability in hearing critical job tasks (Semeraro, 2015; Soli, Amano-Kusumoto *et al.*, 2018,). This magnifies the importance of assessing for this ability in work assessments.

According to the systematic analysis for the global burden of disease (Vos *et al.*, 2017) hearing loss (HL) was ranked among the five leading causes of years lived with disability (YLD). This means more than half a billion people worldwide are afflicted with HL (Wilson *et al.*, 2017). In the United Kingdom (UK) alone, at least 4.4 million affected with HL are of working age (Hearing Link, 2018). A study reported that approximately 32% of the Swedish working population have hearing problems, tinnitus or a combination of both (Hasson *et al.*, 2011). Add to that the median age of workers which is steadily increasing (currently 39 years of age, five years higher than it was 20 years ago). This translates into a larger number of older individuals in the workforce (Wagner-Hartl, Grossi and Kallus, 2018). In the United States (U.S), from the estimated percentage of adults with HL (17%), 55% are in the workforce (Kooser, 2013), bearing in mind that the numbers are estimates as exact data on hearing impairment is sparse, and the actual numbers are probably higher if we take into account the unreported and undiagnosed cases (Stevens *et al.*, 2013). Although it is difficult to compare statistics between countries, it is inferred from the available data that a significant number of people with hearing impairment are in the workforce.

Impaired hearing abilities adversely affect the employee's workplace function in many ways. A survey done on 54 hearing impaired (HI) workers reported the most challenging situations to be meetings, social functions, and training activities. All three settings involve group communication. The challenge lies in the need to increase listening effort and the possibility of missing networking or casual information swapping opportunities. This in turn leads to higher risk of anxiety and frustration (Punch, 2016). Comparisons between normal hearing (NH) and HI workers found that the HI group were at higher risk for mental distress, fatigue and strain resulting in more sick leaves taken (Kramer, Kapteyn and Houtgast, 2006) This increased risk of absenteeism may affect employee productivity.  Efficiency is also affected as tasks may take longer to complete if communications are misinterpreted. Many job settings require the need to hear alarms or signals. Overall, workplace safety can be affected as an individual with HL may jeopardize themselves and others by failing to hear warning signals or important communications (Kooser, 2013). Occupations involving crucial hearing dependant tasks include emergency services, law enforcement (Colaprete, 2012) and military services. For these occupations, tasks depending on hearing are more crucial than others and lack of hearing may be detrimental to the individual and co-workers. Ironically, these same occupations tend to exist in environments that are hazardous to hearing (Grantham, 2012; Le Prell and Clavier, 2017). Studies support the existence of a relationship between noise exposure and workplace injuries (Estill *et al.*, 2017). To avoid hearing-related work injuries, an employer may require the employee to undergo periodical hearing assessments designed to check their hearing-related fitness for work.

### 2.1.1 Causes of hearing loss in the workplace

Hearing loss in the workplace can be due to occupational or non-occupational causes. Non occupational HL may be a pre-existing condition or acquired throughout the duration of employment from causes unrelated to the occupation. Figure 2.1 summarizes the causes of HL in the workplace. Occupational causes are either due to noise exposure, chemical exposure or a synergistic effect of exposure to both noise and chemical ototoxins (Sataloff and Sataloff, 2006).

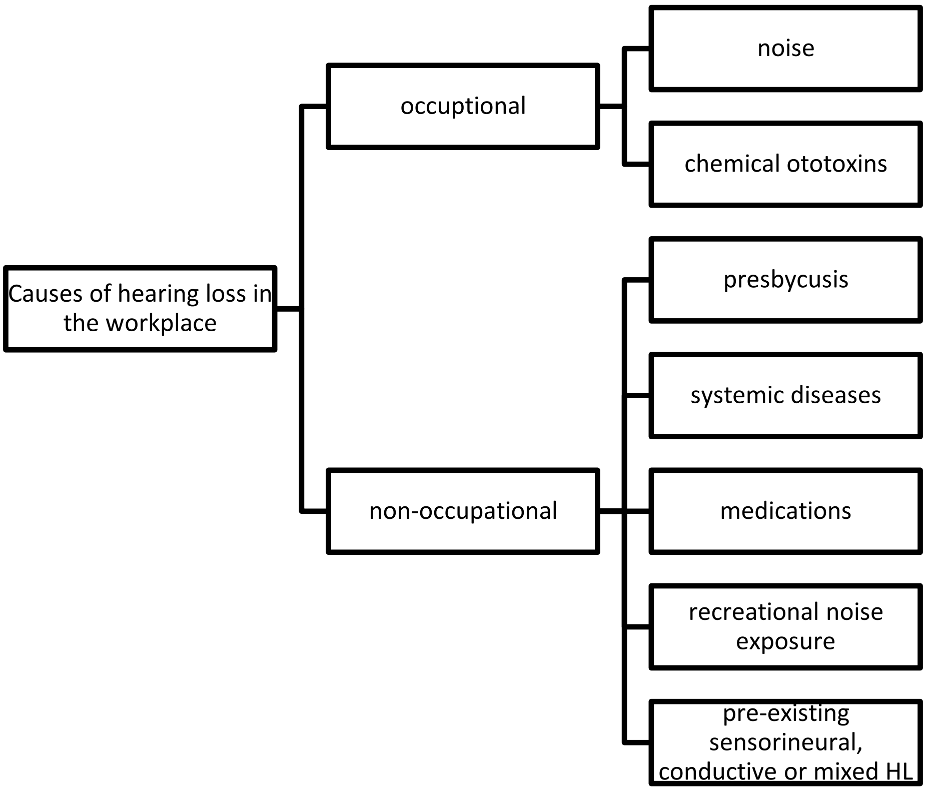

Causes of hearing loss in the workplace
occuptional
noise
chemical ototoxins
non-occupational
presbycusis
systemic diseases
medications
recreational noise exposure
pre-existing sensorineural, conductive or mixed HL

***Figure 2.1*** *Causes of hearing loss in the workplace*

### 2.1.2 Occupational hearing loss

Occupational hearing loss (OHL) is a loss that develops due to a work-related causative factor. It is estimated that 16 -22% of disabling HL is attributed to occupational noise on a global level (Bickenbach, 2011; Huddle *et al.*, 2017). Disabling HL as defined by the World Health organization (WHO) as “a loss greater than 40 dB in the better hearing ear in adults (15 years or older) and greater than 30 dB in the better hearing ear in children (0 to 14 years)” and it is the second most common occupational health problem, after workplace injuries.

Hazards inducing OHL are noise and chemical ototoxins, with high noise exposure being the more common factor (Masterson *et al.*, 2016). Chemicals known to be ototoxic include industrial solvents such as toluene, styrene and carbon disulphide (Sataloff and Sataloff, 2006). The resultant damage is a sensorineural deficit. Noise induced hearing loss (NIHL) is a sensorineural hearing loss (SNHL) that occurs due to chronic exposure to certain amounts of continuous or interrupted noise for a certain period daily. Intense Impulse noise can also affect hearing by causing acoustic trauma (Haboosheh and Brown, 2012) . The effects of impulse noise have been found to be more acute, while continuous noises tend to result in a more slowly developing HL (Lie *et al.*, 2016). The criteria for diagnosis of occupational NIHL differ slightly across regions. In

Saudi Arabia, NIHL is considered an occupational disease if the employee is subjected to a noise level higher than 85 dB for no less than eight hours daily, over a period of no less than five years (General Organisation for Retirement, 2004).

Exposure to harmful levels of noise leads to cochlear damage, starting with the outer hair cells, which translates into damage of the basilar membrane's frequency selectivity (Jansen *et al.*, 2014) and progresses to the inner hair cells and their auditory nerve synapses (Le Prell, 2019). This resulting sensorineural damage affects intelligibility of speech in noise by decreasing speech audibility, an effect which is more pronounced in higher frequencies, and by causing distortion of the speech which is a result of decreased spectral and temporal processing selectivity (Plomp, 1986).

### 2.1.3 Auditory fitness for duty

Assessment of fitness for work is the evaluation of a worker's capacity to work efficiently and competently without risk to their own or others' health and safety (Serra *et al.*, 2007). It is usually assessed as part of the recruitment process, and either periodically after that, depending on the nature of the job, or if there is a change in work condition, location, or employee's health status. Fitness for duty (FFD) testing plays a critical role in the employment status in a variety of settings such as military, law enforcement and firefighting (Zumbo, 2016). Therefore, it should not depend solely on the results of a medical diagnostic test. Multiple factors including specific job requirements, safety issues, legal aspects, employee experience and available skillsets must be taken into consideration. Unlike screening assessments, FFD assessments should be tailored and individualistic while at the same time being cost-effective (Serra *et al.*, 2007)**.**

Fitness for duty tests are high-stakes tests. The consequences of invalid, insensitive, or poorly predictive tests is detrimental to the livelihood of employees (National Academies of Sciences and Medicine, 2019).There has been a growing trend in the past 20 years in the approach to FFD assessments to revisit job-related employment standards and incorporate functional job analyses.

Given the magnitude of HL in the workplace and its possible ramifications, many employers, particularly in critical fast-response jobs, will require assessing their employees' AFFD. Auditory fitness for duty is a sub-section of the wider construct, fitness for duty (FFD). Auditory fitness for duty is the possession of sufficient hearing abilities for effective and safe job performance (Tufts et al., 2009). Occupations requiring AFFD mandate the employee be able to perform hearing critical tasks. If a task requires a specified level of precision and depends on the sense of hearing alone it is a hearing critical task (HCT) (Laroche *et al.*, 2003). This signifies that decreased ability or inability to perform the task successfully will have unfavourable consequences on the effectiveness and/or safety of the employee and /or co-workers (Semeraro *et al.*, 2015). An example of this is a paramedic in an ambulance required to hear and respond to instructions in the background of siren noise. This task depends critically on hearing, requires a specified level of hearing SiN intelligibly, and failure to perform successfully may jeopardize lives, thus making it an HCT.

Assessing for AFFD differs from diagnostic hearing assessment in that it should evaluate specific elements of hearing needed to function in specific work environments to ensure having the minimum required hearing abilities to perform job tasks efficiently and safely (Soli *et al.*, 2018). Often, the standards used by institutes are approached from a purely medical diagnostic angle , blurring the distinctions between diagnostic tests and  AFFD tests (Laroche *et al.*, 2011).

Functional hearing abilities are needed to listen and interact in everyday complex hearing environments. Functional hearing includes the ability to detect and localize sounds, the ability to recognize speech, and to understand it (National Academies of Sciences, Engineering, and Medicine, 2019). The abilities needed for functional hearing are not restricted to hearing acuity but involve multiple processes such as temporal processing and cognitive abilities. While AFFD protocols can be comprised of a battery of tests, sometimes one test is required by the employer. In this case, the most representative test must be chosen.

Pure-tone audiometry (PTA) is still the mainstay of AFFD assessment worldwide and while it has been recognized over the years as a vital diagnostic tool for audiological conditions, its boundaries are the monoaural audibility threshold of the peripheral hearing system in quiet. Pure-tone audiometry is a poor predictor of functional real-world hearing and SiN intelligibility (Soli, Amano-Kusumoto *et al*. 2018), and may not be a representative measure of SNHL in noise. Many Jobs requiring AFFD take place in noisy environments, where noise levels usually exceed 70 dB(A), and while sound detection in quiet is important, it is probably not the primary ability affected in performing HCTs.

A study identifying the most important auditory tasks deemed critical to missions of the British infantry and combat-support personal was carried out by Semeraro *et al.* (2015). Prioritization was done according to task frequency and consequences of poor task performance. This resulted in the identification of seven speech communication tasks, one localization and one detection task. This indicates the commonality and importance of speech tasks in military environments.

Hearing and understanding speech in noise are more requisite abilities for performing HCTs in noisy environments. Although some studies have shown correlation between PTA scores and results of speech intelligibility tests (Smoorenburg, 1992; Picard *et al.*, 1999; Jansen *et al.*, 2013), many have demonstrated the PTA's inability to accurately predict the variation in speech in noise intelligibility performance between individuals (Smoorenburg, 1992; Killion *et al.*, 2004; Wilson, 2011;Vermiglio *et al.*, 2012; Brungart *et al.*, 2017). A systematic review of the strengths and limitations of the PTA (Musiek *et al.*, 2017) reported multiple studies highlighting the ineffectiveness of PTA at aiding diagnosis of real-world signal perception deficits or disorders of the central auditory system. One of the reviewed studies compared veterans with and without blast related injuries. Both groups presented with normal PTA thresholds, but the blast exposed group exhibited notable deficits in SiN recognition, temporal resolution and binaural integration (Gallun *et al.*, 2012). Mild traumatic brain injury (MTBI) is a consequence of blast injuries, falls and other causes all which military employees are at risk for. In a study comparing between SiN performance in listeners with MTBI and a matched control group, significant differences in performance were found between individuals with subjective difficulties in SiN interpretation and those without (Hoover, Souza and Gallun, 2017). United states Army reports from superior officers and military hearing specialists noted that soldiers being classified as H3, detailed in Table 2.1, by PTA had wide variation in their abilities to perform mission tasks, the consequences of which are either negative to the employee who is unnecessarily removed from post or to the institution if someone unfit stays and is a source of jeopardization (Brungart *et al.*, 2017). Only upon reaching a classification of H3 would SiN testing be incorporated into the remaining assessment.

***Table 2.1*** *U. S Army hearing profiles, adapted from (Brungart et al., 2017)*

| Hearing category | PTA average at 500,1000 and 2000 Hz | PTA level of individual frequencies | PTA level at 4000 Hz |
|---|---|---|---|
| H1 | not exceeding 25 dB | not exceeding 30 dB | not exceeding 30 dB |
| H2 | not exceeding 30 dB | not exceeding 35 dB | not exceeding 55 dB *or* not exceeding 35 dB in the better ear in the presence of a poor ear |
| H3 | SRT not exceeding 30 dB in the better ear (aide or unaided) | | |
| H4 | Hearing loss exceeding the defined level of H3 | | |

Sheffield *et al.* (2015) conducted a dismounted combat paintball simulation task using immersive simulated HL systems simulating varying degrees of HL. They found that in a condition not requiring communication, apart from severe HL, hearing impairment did not affect performance much. There was a decrease in lethality of individuals proportional to decreased hearing due to change in behavioural survival tactics, however, task performance of survivability depending on sound detection and localization was otherwise unaffected. This demonstrates the importance of task analysis and awareness of job requirements. It also highlights the importance of adequate AFFD assessment as even those with moderate hearing loss were able to perform the task. Another case report examined an aircraft crewmember of the U.S army who was deemed unfit by the AFFD testing protocol but was then granted a waiver after undergoing functional tests reflective of his duties (speech intelligibility testing in a number of simulated work conditions in the aircraft with and without a helmet), highlighting the importance of measuring what is required on the job; in this case hearing speech in noise, as opposed to the deficient protocols in place (Casto and Cho, 2013).

Inadequate standards may result in legal repercussions, a natural consequence for an employee whose livelihood is affected while they perceive themselves to perform adequate to their peers. There has been a recent shift toward changing the evaluation criteria for AFFD assessment and a broader acceptance of functional assessment (Brammer and Laroche, 2012). Several organisations have demonstrated efforts to revise their hearing standards and implement more effective AFFD standards. These include the Royal Canadian Mounted Police (RCMP), California's correctional officers, the Canadian Coast Guard and Department of Fisheries and Oceans (DFO), Ontario's Constable Selection System, the UK infantry and the US army. (Laroche *et al.*, 2011; Vaillancourt *et al.*, 2011; Semeraro *et al.*, 2015; Tufts *et al.*, 2018; Giguère *et al.*, 2019).

Speech in noise tests have been proposed as a more functional measure of hearing speech in noisy work environments. The Canadian DFO utilized the scores from the hearing in noise test (HINT) in stationary speech-spectrum noise (SSSN), in a model designed to predict the individual's performance in the real-world noise environments of the DFO, considering communication distances, vocal effort, repetition of commands and minimum % intelligibility for ensured task performance (Giguère *et al.*, 2008). Development of a portable test protocol (using a tablet and headphones) that assesses all required functional abilities for US military is underway. This includes a PTA, a test for recognition of military sounds in noise; the MILSINT, and a SiN test (Tufts *et al.*, 2018). From a large test battery, a SiN test and a test of binaural masking level difference were found to be the most sensitive to suprathreshold hearing, and suggested as predictive tools for functional hearing assessment (Phatak *et al.*, 2019; Grant *et al.*, 2021).

Many organizations worldwide rely on PTA for AFFD assessments, though studies have shown that tests of a more functional nature are also required (Tufts, Vasil and Briggs, 2009; Soli, Roeser and Vaillancourt, 2019). To date, US service members are only required to do an annual audiometric screen with no further testing required if hearing is within accepted limits. This includes service members who have been exposed to blasts or harmful noise, acknowledging our current knowledge of the suprathreshold effects within these individuals (Grant *et al.*, 2021)

While some institutions might include additional tests if a specific degree of HL is found based on PTA results such as the U.S army, PTA is the only AFFD assessment measure in Saudi Arabia in governmental, military and presumably most private sectors (General Organisation for Retirement, 2004; Al-Omari *et al.*, 2018). Due to this insufficiency in functional occupational assessment and paucity of functional assessment tools, it was decided to focus the PhD around AFFD and develop an Arabic measure of AFFD to be researched for this purpose.

A Cochrane review on 11 studies that were randomized controlled trials (RCTs) or interrupted timeseries studies evaluating the effectiveness of general or job related pre-employment screens concluded that there was low quality evidence that job specific pre-employments screens may have a positive effect on FFD, supporting only job-specific testing (Schaafsma *et al.*, 2016). This was based mainly on a study involving the Netherlands army, which randomized the old army examination system (PULHEEMS) with a newer workload capability based model (BMEKL) (De Raad and Redekop, 2005; De Raad, Nijhuis and Willems, 2005). None of the included studies were on AFFD. Although there is a wide body of literature on FFD, the available studies on AFFD are quite few since the concept is somewhat recent. Also, none of the studies developing or exploring AFFD measures have reached the point of RCTs.

The goal in development of AFFD standards and protocols is to decide on a criterion level in a suitable hearing assessment that equates to the minimum performance level required for safe and effective job performance. This initially requires conducting a detailed job task analysis. From this analysis a compilation of HCTs is identified. They may also be grouped based on frequency and criticality. This step is done jointly by specialists in hearing sciences and subject matter experts (SMEs) on the job. This is followed by identification and analysis of work noise environments. The required level of auditory performance in each noise environment should be objectively established. Then, appropriate hearing assessments reflecting required tasks in their respective noise environments should be selected accordingly. This process was used to establish AFFD standards for Ontario's constables (Giguère *et al.*, 2019) by building on the work done by Soli *et al.* (2018). A smaller scale study was done to develop an AFFD measure for UK military infantry personnel, where analysis of mission critical tasks was carried out and a suitable SiN test; the British- English Coordinate Response Measure test (CRM), was developed and evaluated (Bevis *et al.*, 2014; Semeraro *et al.*, 2015, 2017).

There is a need to differentiate between performance standards and cut off scores. The cut off score is the point below which the test taker fails to meet the minimal adequate level of performance, known as the performance standard. The performance standard is a conceptual required level of competence and the cut off score is the operational level of competence representing the conceptual level. When things are clarified accordingly, the cut-off levels are more defensible. Many FFD measures are somewhat arbitrary in the sense that they do not seem to reflect the required level of performance in the job tasks (De Raad, Nijhuis and Willems, 2005; Serra *et al.,* 2007). It is also important to address the point that accepting cut off scores is a balance between human, financial and administrative factors. There is a very fine balance between inclusivity and high efficiency in an organisation.

Given the complexity of fully developing AFFD standards, the full development procedure is beyond the scope of this PhD. However, the lack of tools for studying AFFD in an Arabic Saudi population, and the shortcomings of PTA in providing a holistic picture of important aspects of a person's functional hearing such as everyday speech perception (Williams-Sanchez *et al.*, 2014), motivated the development of a tool that tests a critical hearing function; recognizing speech in noise, to be further explored as a suitable measure for AFFD studies. To better understand assessments of AFFD, the mechanism of speech perception in noise is reviewed in the following section.

## 2.2 Speech perception in noise

### 2.2.1 How is target speech recognized in the presence of background noise?

The mechanisms involved in speech perception in noise; coined the cocktail party effect by Cherry (1953), consist of bottom-up cues entering the peripheral auditory system, and higher level central controls acting from the top-down. As illustrated in Figure 2.2, it is not a one-way process. Rather, bottom-up salience of incoming spectro-temporal, visual and localization cues influences top down controls, and top down subcortical, cortical and attentional controls guide the selection and attentiveness to bottom up salient features (Shinn-Cunningham, Best and Lee, 2017). As convenient as it would be to have the mechanisms of object formation and selection occur in sequence, they are somewhat simultaneous and continuously subject to feedback from one another, in addition to constant influences of dynamically changing stream inputs and attention.

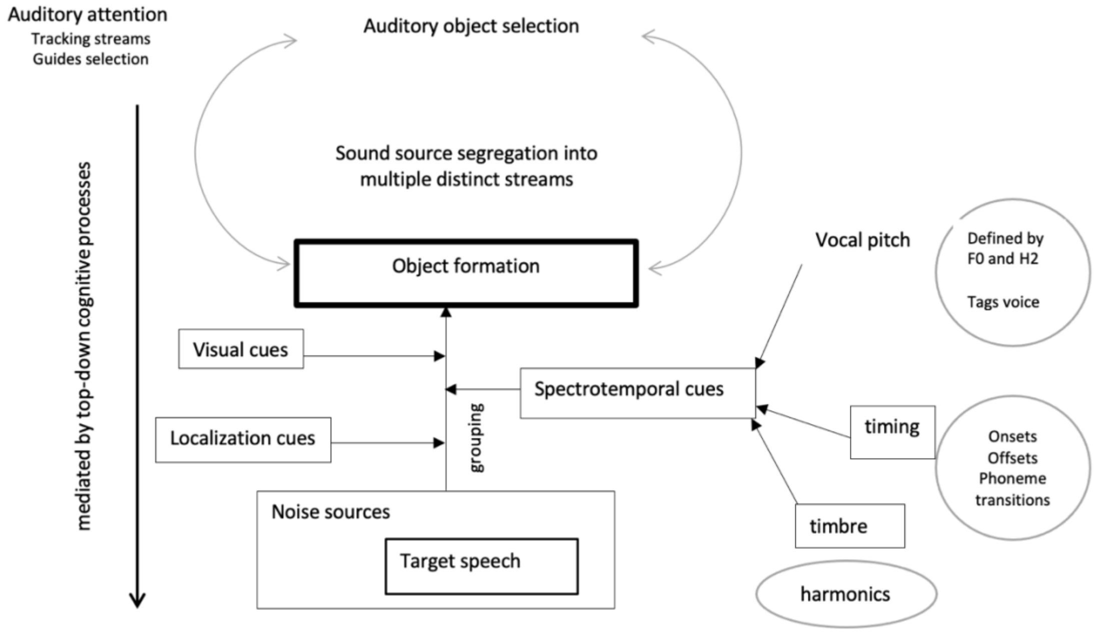


***Figure 2.2*** *Simplified depiction of the process of listening to target speech in noise*

To understand the roles of central processes in SiN perception, the sensory auditory pathway is overviewed in Figure 2.3. The cochlear nuclei of the superior medulla receive unilateral sensory inputs from the cochlea via the cochlear portion of the vestibulocochlear nerve. These inputs travel up to the superior olivary complex (SOC) in the pons. Integration and interpretation of bilateral inputs occurs at this level (Christov, Nelson and Gluth, 2018). In the inferior colliculus (IC) of the midbrain, auditory information is integrated and projected to the auditory cortex (in the temporal lobe) via the medial geniculate nucleus of the thalamus. The IC is also the point of reception for extra-auditory information. The role of the brainstem and cortex is covered in section 2.2.3.

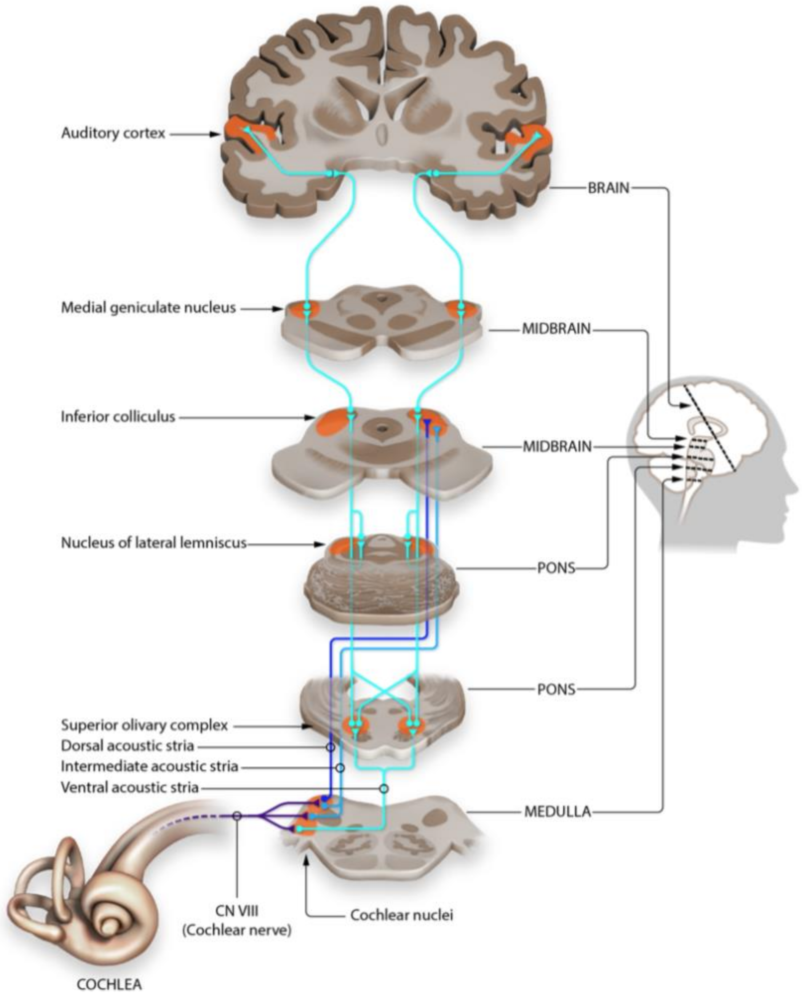


***Figure 2.3*** *Human neural auditory pathway by Jonathan Peele (2018) (License: CC BY 4.0)*

### 2.2.2 Peripheral cues

The perception of speech starts with physical acoustic cues originating from a pool of sound sources that are processed, parsed and streamed on peripheral and cognitive levels (Bregman, 2001). Speech is very rich in spectro-temporal cues to the point of redundancy, which is why the system can extract speech from noise even under adverse conditions.

Audibility is essential to speech intelligibility in noise (Humes, 2007; Sanchez-Lopez *et al.*, 2021; Wasiuk *et al.*, 2022). On a suprathreshold level, frequency selectivity is a critical function for speech intelligibility in noise. The cochlea acts as a bank of bandpass filters, each filter with a different bandwidth and central frequency. Situated at the apex are narrower filters tuned for detection of lower frequencies, hence higher selectivity. This selectivity is paramount for intelligibility of SiN as it allows the auditory system to differentiate components of complex auditory stimuli (De Sousa *et al.*, 2020). The nature of spread of sound sources on the frequency map of the basilar membrane of the cochlea results in overlap of the spectral source representations. (McDermott, 2009).

Temporal cues are the auditory signal changes observed over time. They can be divided into two groups; illustrated in Figure 2.4: the envelope, which corresponds to the slow fluctuations of amplitude over time (modulating frequency) and the temporal fine structure (TFS) which corresponds to the rapid oscillations with rate close to the centre frequency of the band. Experiments by Hopkins and Moore (2010) suggest that TFS has a significant role in formant perception (Gnansia *et al.*, 2009) and hearing SiN (Moore, 2013) and may be critical for perceptual segregation, while the envelope also influences speech intelligibility (Swaminathan *et al.*, 2016).

Sensorineural hearing loss associated with aging and NIHL reduces the ability to glimpse the signals amidst the noise's temporal and spectral fluctuations, due to loss of frequency fine-tuning, a smearing or distortion caused by flatter psychophysical tuning curves. A similar mismatch between the TFS information and the place corresponding to it on the basilar membrane reduces the ability to process TFS information (Moore, 2008). This has a detrimental impact on SiN intelligibility. Baer and Moore (1994) demonstrated the adverse effect temporal smearing and decreased selectivity had on the intelligibility of speech in noise.

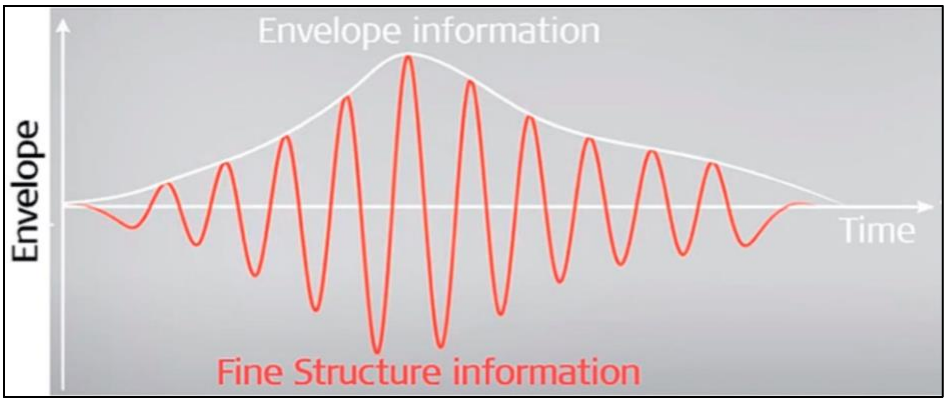


***Figure 2.4*** *Simple illustration of the envelope and fine structure components of a sound signal (image courtesy of MED-EL) (Dhanasingh and Hochmair, 2021).*

Other cues to be considered are binaural cues. These include inter-aural time difference (ITD) which corresponds to the delay in time between the sound reaching the two ears, inter-aural level difference (ILD) which is the difference in loudness between the sound reaching the two ears, and binaural coherence (Bronkhorst, 2015). These differences are related to head shadow effect and differ across frequencies. Considering frequency wavelengths, we find that ILD is more pronounced at high frequencies and insignificant below 500Hz. As for ITD's they are more beneficial at low frequencies and ambiguous at higher frequencies (Moore, 2013). Inter-aural level differences have a dominant effect on grouping while the effect of ITDs is less pronounced (Bronkhorst, 2015). The binaural system aids spatial selectivity, which in turn aids higher speech intelligibility (Andreeva, 2018). Spatial cues are very effective in speech segregation (Bronkhorst, 2015), and depending on the focus of the listeners attention, may improve object formation and selection (Ihlefeld and Shinn-Cunningham, 2008).

With the exception of amplitude modulation, many of these cues have a weak influence on local grouping on their own, however, abstractly speaking when added together, they correlate through their regularities to give the perception of sound from one source (Shinn-Cunningham, Best and Lee, 2017). More important than the spectro-temporal cues in isolation, are the principles that perceptually organize them. The determinants of simultaneous grouping are similarity, smooth continuation of the changes in the physical components, common synchronous onsets and offsets, and the continuity effect which is the phenomenon that makes a sound seem continuous even when it is intermittently or temporarily masked by another sound (Moore, 2013).

All physical cues are adapted and grouped by the principles of simultaneous grouping into auditory objects. Each auditory object is then streamed separately through time. Streaming is governed by the principles of sequential grouping (Bregman, 2001). These cues activate the recognition process at different times depending on cue nature so the relative timing of the cues plays a role (Mattys *et al.*, 2012). Cues affecting sequential grouping are acoustic transitions, ITD's and judgements of temporal order (Moore, 2013). One stream is then selected and attended to. This is where higher processes and attentional roles become evident.

### 2.2.3 Beyond the periphery

The brainstem assists interpretation of spectro-temporal cues required for SiN perception. Strong neural brainstem responses reflecting encoding of vocal pitch have been documented by EEG (Bidelman and Momtaz, 2021). The pattern detection and information extraction from vocal pitch cues is facilitated by rostral (cortical and upper brainstem) and caudal (lower brainstem) efferent innervations (Maruthy, Kumar and Gnanateja, 2017). The medial olivocochlear (MOC) bundle; part of the SOC, modulates the response of the outer hair cells (OHCs) to sound by providing efferent feedback (Smith and Cone, 2021).

Binaural squelch is a cognitive process that combines binaural information to enhance speech comprehension. It is an important localization cue for segregation. Fusion begins at the brainstem level in (SOC), where dichotic spectral, amplitude and temporal information is integrated, giving dimensional information, and suppressing noise to improve the speech signal. This in turn aids parsing auditory objects and improves speech intelligibility in noise (Glyde *et al.*, 2011; Bipin Kishore, 2020).

The study of auditory object formation stemmed from theories of visual attention. Similarities that control the perception of both vision and audition propound that they operate under the same attentional neural mechanisms. Studying the auditory domain is more challenging due to the perceptual nature of the cues forming the objects, and the temporal nature of streaming (Shinn-Cunningham, Best and Lee, 2017). Research on the specific function of attention is hindered by the heterarchical interactions between bottom-up salience of sensory stimuli and competing streams affecting top-down attention. Rapid attentional shifts serve to further complicate matters (Bronkhorst, 2015).

Sounds are organized in the auditory memory by streams (Bregman, 2001; Sussman, 2017).The focus of attention in streaming is strongly influenced by peripheral syllabic components such as location and timbre. Different cues activate different areas in the cortex. Spatial attention is associated with higher stimulation in the dorsal frontoparietal network, and pitch-based attention is associated with greater stimulation of the auditory processing areas of the inferior frontal gyrus (Shinn-Cunningham, Best and Lee, 2017).  Attention can parse background task irrelevant sounds that were not automatically parsed by pre-attentive processes. This is supported by Wild et al. (2012) who used fMRI to demonstrate that unattended speech was processed and that activity was elevated in attentional centers when attending to degraded speech. Information of unattended sounds is held in the memory to be accessed by task-specific and automatic processes.

The auditory cortex is actively involved in stream segregation. Functional studies show higher-order auditory cortical activity reflective of attending to a speaker to be higher in fidelity than the activity to unattended objects while primary cortical areas display activity to both attended and unattended objects (Puvvada and Simon, 2017). Cortical slow neural oscillations both support and are guided by the structure of the attended sound. This evidences the crucial role of cortical processes (Grant *et al.*, 2021) in auditory situational awareness and attendance to speech. Modulation of cortical evoked responses to sound by selective attention is subject to individual differences (Choi *et al.*, 2014). This highlights the sensory-cognitive relationship in SiN perception.

## 2.3 Work-related hearing assessments

Occupational hearing assessments serve two purposes, hearing screening and assessment for AFFD. Hearing screening is usually done as part of a hearing conservation programme (HCP). Hearing conservation programmes provide awareness through education to decrease risks of noisy environments in addition to providing protection and screening. The basic elements of occupational HCPs are noise exposure control, employee education about the importance of hearing conservation, providing hearing protection devices (HPDs), periodic hearing screening assessments and diligent monitoring of progression through record keeping (Occupational Safety and Health Administration, 2002). Timing is key in screening tests. Early testing is imperative to detect early changes and allow for early interventions to avoid progression to disability (Stenfelt *et al.*, 2011). The second purpose is assessment of hearing capabilities for AFFD purposes.

### 2.3.1 Auditory fitness for duty assessments

Most jobs will encompass tasks involving comprehending speech or other signals in adverse listening conditions. With our knowledge of the auditory system and speech perception in noise, the complexity of the mechanism and the involvement of peripheral and central integrated processes are apparent. These should be measured accordingly. Surprisingly, despite the widespread knowledge of the PTAs shortcoming as a valid predictor of functional real-world hearing, it is still the main indicator of AFFD with a few sectors including SiN testing as a secondary stage of assessment in employees not meeting the criteria of the PTA (Soli, Amano-Kusumoto, *et al.*, 2018). Most conventional hearing testing involves PTA only, which is a measure of peripheral hearing sensitivity. Most jobs requiring adequate AFFD are not limited to detection of simple sounds in quiet, the construct measured by PTA. It is vital to consider the equity of AFFD assessment standards as they are considered unfair if they do not fulfil the following criteria:

- They are based on detailed job analyses
- They are valid indicators of the HCTs derived from the job analyses (Payne and Harvey, 2010).

They must also take into account the amount of variability in performance related to variability among humans, while maintaining test-retest reliability (Dobie and Van Hemel, 2005).

Speech in noise tests are most likely better predictors of communicating in noise. The limitation of SiN testing lies in the difficulty of test standardization. Unlike PTA, there are different speaker properties, background noises and speech materials, in a variety of languages. This variation affects the sensitivity of SiN tests which is why careful selection of the speech material and background noise is of the utmost importance (Jansen *et al.*, 2014).

### 2.3.2 The relationship between SiN tests and PTA

The relationship between PTA and SiN tests is debated across the literature. Appendix A lists a number of studies looking at the relationship between hearing sensitivity and SiN test performance, discussed in this section. Some studies used additional measures such as speech in quiet or cognitive assessments.

Merten *et al.* (2022) found hearing sensitivity to be the strongest determinant of SiN perception in adults above 30 years of age. They also found crystallized intelligence and executive functions to have stronger associations with SiN performance, while working and long-term memory had much smaller independent effects. Pure-tone audiometry was the best predictor for unaided SiN performance in elderly hearing-impaired listeners in Gieseler *et al.*'s (2017) study. Speech recognition in noise was influenced mainly by high-frequency hearing loss, a strong predictor, but also to some extent by age and mid-frequency hearing (Barrenäs and Wikström, 2000). Schoof and Rosen (2014) found PTA to be the strongest predictor ($r^2 = 0.32$, $p = 0.001$) in normal hearing adults across age groups in babble noise, but not as strong a predictor of SiN in SSSN.

Other studies found a poor relationship between SRT and PTA ( Vermiglio *et al.*, 2012). However, upon reassessment, they found a stronger relationship between PTA and SRT in people with partial audibility of SRT (Vermiglio *et al.*, 2020). Stenbäck, Hällgren and Larsby (2016) reported no association between PTA and SiN scores in NH young adults, only in older adults. A structural model used to evaluate factors contributing to SiN understanding found no significant prediction of peripheral hearing measured by PTA (Anderson *et al.*, 2013). Some studies found correlations with PTA frequencies above 1KHz more predictive of SiN recognition (Smoorenburg, 1992). Holmes and Griffiths (2019) found high frequency PTA moderately correlated with SiN recognition in normal hearing individuals explaining 15% of variance. Phatak *et al.*(2019) found the highest correlations between the composite scores of a functional hearing test battery and PTA scores of participants with mild to moderate hearing loss at 0.5, 1, and 2 kHz accounting for 33% of the variance in SiN scores. They suggested supplemental SiN testing with PTA for comprehensive hearing assessment. Overall, there is no clear agreement on the extent of the relationship and predictive ability of the PTA to SiN performance. Clearly, it is not a precise enough predictor of individual ability to recognize speech in noise. As discussed in section 2.2, the complex grouping of cues needed for perceiving speech in noise involves muti-level processes not utilized in detection of pure-tones.

### 2.3.3 Speech intelligibility testing for occupational assessments

Speech intelligibility is the ability to correctly identify a sentence or words from a larger selection of sentences or words without necessarily completely understanding the identified content. It is equivalent to speech identification or recognition. In the hierarchy of auditory skills it is preceded by speech detection and discrimination and superseded by speech comprehension (Erber, 1982 cited in National Academies of Sciences, Engineering and Medicine , 2019). To assess speech intelligibility, a psychophysical method is required. This method expresses the relationship between the acoustic stimulus presented, and the individual's perception of the stimulus. The responses to the stimuli tested are represented as distributed points on a psychometric function (PF) (Kingdom and Prins, 2010). Psychometric functions are explained in section 3.3.3**.** Several tests have been developed or used to assess SiN for occupational/AFFD purposes. They are listed in Table 2.2.

A search for SiN tests used for occupational assessment purposes was conducted in DelphiS which uses EBSCOhost research platform. It covers 350+ databases including Medline, PsycINFO, CINAHL, Science direct, EMBASE and Medcom. The search was done using the following Boolean phrases:

- 'Speech in noise test' AND 'occupational' generating 14 results
- 'Speech in noise test' AND 'auditory fitness for duty' generating six results

After excluding irrelevant results and grouping studies researching the same test, the search yielded six tests, used, or developed with the intent of use for occupational purposes. None of them were available in Arabic. The tests are briefly described in Table 2.2 in terms of available languages, content and noise background.

***Table 2.2*** *Speech in Noise tests developed or used for occupational assessment purposes*

| **Test** | **languages** | **Content** | **Noise** | **Other** |
|---|---|---|---|---|
| Hearing in noise test (HINT) (Laroche *et al.*, 2003) | 13 languages including American English | Sentences | Headphones or 8 speakers, 1 playing signal and 7 playing steady state speech shaped noise | Adapted from BKB<br>Male voice<br>Adaptive signal to noise ratio (SNR) |

| **Test** | **languages** | **Content** | **Noise** | **Other** |
|---|---|---|---|---|
| Speech Recognition in noise test (SPRINT)<br>(Brungart, Sheffield and Kubli, 2014) | American English | Monosyllabic words<br>100 words (modified from the original 200-word lists) | Multi-talker babble via earphones | Fixed SNR<br>Used by the U.S army |
| Coordinate Response Measure (CRM) test<br>(Semeraro *et al.*, 2017) | American and British English | Sentences | Steady state speech shaped noise via headphones or loudspeakers | Adaptive SNR<br>Developed for UK infantry |
| Digit Triplet Test (DTT)<br>(Smits *et al.*, 2004) | Eight languages including English and German | Digit triplets e.g., one-two-six | Steady state speech shaped noise via headphones or through the phone | Adaptive SNR<br>Used mainly for screening purposes |
| Occupational Earcheck (OEC) (Ellis et. al cited: Sheikh Rashid and Dreschler, 2018) | Dutch | CVC syllables | Long term average speech spectrum (LTASS) stationary noise through headphones | Adaptive SNR<br>Mainly developed for internet-based screening purposes |
| Clinical version of the modified rhyme test (MRT)(Brungart, Makashay and Sheffield, 2021) | American English | Two 80-word lists | Speech-shaped noise presented binaurally through headphones | Binaural<br>Two fixed SNRs |

The search for Arabic validated material was conducted numerously throughout the PhD period to disclose any tests made available. Regarding Arabic SiN tests for adults, recently an Egyptian Arabic version of the Quick SiN test was developed (Elrifaey *et al.*, 2021). However, this test is not of much use because of the accent difference which is not representative of the country's dialects and would adversely affect listener performance.  An Arabic version of the DTT was presented in a conference paper at the 2016 Annual Meeting of the German Society for Audiology Literature with no explanation or reference to the test development details (Koifman *et al.*, 2016). The author of this paper was contacted for further details but did not reply. The development of an Arabic version of the matrix SiN test was presented at the 4th conference of the Advanced Arab Academy of Audiovestibulogy, but there are no publications on either test to date. Currently, a multi-site study is on-going to validate the Arabic paediatric version of the Matrix test and the funding company, Cochlear, have confirmed the adult Arabic version has yet to be published (Yalkawi, 2022).

## 2.4 Factors influencing SiN test performance

The main goal of this research was to develop and research a measure of AFFD relevant to occupations of a specific nature, that is fast-response jobs involving listening to and executing commands, such as the police force or military services. The Royal Saudi Air Defense Forces (RSADF) expressed interest in a test that would identify recruits having sufficient hearing to perform tasks upon entering the military and throughout their career.

In this section, factors affecting recognition and comprehension of speech in noise as a requisite ability for AFFD are addressed, with the military in mind. Listening to speech in a noisy environment is a complex function with many contributing factors including spectro-temporal and spatial cues, audibility, vocal characteristics, linguistic and syntactic features, effects of reverberation, and higher cognitive processes including attention and working memory. This section will discuss the most important aspects affecting the ability to hear SiN and take into consideration factors important for AFFD.

### 2.4.1 Noise environment

Most everyday environments contain background noise that may mask speech and other sound signals affecting intelligibility. Masking refers to the imperceptibleness of a signal due to another present sound. There are two types of masking noise, energetic maskers which interfere with the target signal by physically degrading it spectrally and temporally, and informational maskers that contain information acoustically or linguistically resemblant to that of the target, hindering object formation and/or segregation of the target signal from the masking noise through higher-level masking processes (Mattys *et al.*, 2012; Goossens *et al.*, 2017).

Stationary noise provides more energetic masking than fluctuating noise; depriving the listener of "glimpses" to attend to the target. Some fluctuating maskers are more difficult than stationary maskers due to the summative effect of the energetic and strong informational component such as two-talker- (van Engen, 2012; Schoof and Rosen, 2014) or four-talker babble (Cainer, James and Rajan, 2008). Although multi-talker babble (8 and above) is considered an energetic masker, it is still less disruptive than SSSN due to the amplitude modulations (Theunissen, Swanepoel and Hanekom, 2009). It is also important to consider the nature of the babble and the talker. Native language babble provides stronger informational masking than non-native babble (Mattys *et al.*, 2012). Similarity of masker and target voices also strengthens the informational masking (Brungart, 2001).

Listening can occur in an outdoor environment subject to weather conditions and wide distances or indoor environments containing ambient noises of machines, air conditioners and reverberation. Reverberation is an acoustic phenomenon resulting from the superimposed effect of direct sound and its reflection from room surfaces or objects. It affects speech recognition by interfering with spatial auditory attention (Ruggles, Bharadwaj and Shinn-Cunningham, 2011), in turn affecting unmasking and segregation, especially in elderly or  hearing compromised individuals (Picou, Gordon and Ricketts, 2016; Xia *et al.*, 2018). The distribution of sound sources also plays a role in the SiN perception. Listening to moving sound sources is more challenging than fixed sources (Shinn-Cunningham and Ihlefeld, 2004). Absence of spatial separation creates more difficult listening conditions as spatial release from masking is a very helpful cue in improving hearing in noise (Soli and Wong, 2008a; Coffey, Mogilever and Zatorre, 2017). This effect is more pronounced with energetic maskers.

### 2.4.2 Nature of auditory stimuli

Stimuli used in assessing SiN are either phonemes, words, or sentences. Words are less representative of real-world communication and lack cues present in continuous speech that listeners rely on to extract information from higher order lexical storage (Killion *et al.*, 2004). Words however, are not subject to strong reliance on memory or cognitive skills (Le Prell, 2019). Contextual cues found in sentences help the listener to ascertain the signal with less reliance on acoustic cues, for this reason matrix type tests are more difficult due to absence of these cues.

When the speech is in a non-native language, performance in noise is more adversely affected due to higher level lingual processing limitations (Mattys *et al.*, 2012). Talker familiarity improves performance as opposed to different talkers (Theunissen, Swanepoel and Hanekom, 2009) . Other auditory sources adversely affecting SiN recognition include accented or disfluent speech and communication channel degradation such as telephone filtering (Mattys *et al.*, 2012).

### 2.4.3 Learning effect

Learning in the context of SiN tests is categorized into content and procedural effects. Content learning results from familiarization with the test material (Theodoridis and Schoeny, 1990; Yund and Woods, 2010). Perceptual learning is the improved performance in perceptual tasks resulting from training (Zhang *et al.*, 2019). This effect occurs with all SiN tests and should be overcome quickly and efficiently to reach the stable true performance levels (Schlueter *et al.*, 2016). The learning effect is more problematic in open-set tests, while with closed-set tests, familiarization is faster due to limited material, allowing quicker mitigation of the learning effect by shorter training protocols (Brungart *et al.*, 2022). For matrix type tests, two consecutive trainings consisting of approximately 20-30 trial sentences are recommended (Hagerman and Kinnefors, 1995, Schlueter *et al.*, 2016, Willberg *et al.*, 2020). Even for long-term learning, the amount of learning in a matrix type test plateaus after a certain time (Willberg *et al.*, 2020). This is reassuring if a test is designed to test individuals periodically.

Whether intra- and inter- session learning effects are greater in HI impaired individuals is debatable in the literature. Studies found the effect to be larger in HI individuals, requiring more training to be overcome (Hagerman and Kinnefors, 1995) while other studies found the effect to be smaller than their NH counterparts (Wagener and Brand, 2005a; Schlueter *et al.*, 2016). Learning effects are more substantial with fluctuating noise backgrounds than stationary backgrounds and may require more training to be overcome (Rhebergen, Versfeld and Dreschler, 2008; van Engen, 2012).

### 2.4.4 Experience and training

The role of experience in understanding speech in noise is documented in several studies. Air traffic controllers, employees accustomed and trained to communicate in noise, significantly outperformed a normal hearing control group in speech in noise tasks with the largest difference seen in the most difficult listening conditions (Zaballos *et al.*, 2015). Semeraro (2015) observed improved performance in a miliary population compared to a civilian sample but the difference in performance was only evident in task-related simulations, not in generic speech tests.

The effect of training to listen in noise is also documented outside the context of work-related training. Word-based training improved word recognition scores in young and older adults but this effect was not transferred to performance with sentences in noise even when trained words were embedded in the sentence (Burk *et al.*, 2006). Older adults undergoing a 6-month musical training program had improved performance on a word-in-noise recognition task (Zendel *et al.*, 2019). Clearly the training effect appears to be context specific with a learning specificity. A more recent study found training and learning generalized when implemented in an audio-motor training videogame that tapped into multiple attentional foci and functional abilities, requiring participants to shift focus between different psychophysical tasks and providing dynamic closed-loop feed-back. The control group was trained with working memory tasks. Working memory is a cognitive storage system that is discussed in section 2.4.6. Compared with the control group, the video-game group had a 25% improvement in SiN recognition, but the effect was transient and declined with discontinued training (Whitton *et al.*, 2017). This links to the findings discussed above, where training to recognize words and sentences does not transfer to beyond training contexts, while continuous training such as a job requiring frequent listening in noise or musical training improve SiN recognition outcomes.

### 2.4.5 Hearing abilities

Speech comprehension is a vital functional hearing ability for communication (Dubno, 2018). Often at work, communication is required in noisy environments. Hearing is one of the more well-known factors affecting SiN performance and in turn AFFD (Stenbäck, Hällgren and Larsby, 2016; Goossens *et al.*, 2017; Hwang, Kim and Lee, 2017; Ross, Dobri and Schumann, 2020). It links to age and noise exposure among other factors. Auditory handicap is most evident in functional abilities such as hearing speech, usually in some type of form of background noise, and assessment by tones in quiet is not representative of this real-world performance. These abilities play a major role in AFFD performance, and as such must be assessed accordingly (Tufts, Vasil and Briggs, 2009; Soli, Giguère, *et al.*, 2018; Giguère *et al.*, 2019).

### 2.4.6 Age and cognitive abilities

Deterioration of SiN test performance in older age groups may occur even in the absence of an apparent peripheral pathology (Pürner *et al.,* 2022). Effects of auditory aging on SiN performance include decreased hearing sensitivity and dynamic range, declining efficiency of physiological processes such as a decrease in temporal resolution or frequency selectivity, and reduced spatial and cognitive processing (Burk *et al.*, 2006; Glyde *et al.*, 2011; Innes-Brown *et al.*, 2016). Attention, working memory, and motivation all influence behavioral test performance (Anderson and Kraus, 2010). Cognitive performance influences SiN performance in adverse listening conditions (Stenbäck, Hällgren and Larsby, 2016; Ross, Dobri and Schumann, 2020).

Anderson et al (2013) found working memory to be the most influential cognitive function in understanding SiN. Working memory (WM) is a limited capacity system for temporary storage and processing of actively used task-relevant information needed for speech comprehension (Schoof and Rosen, 2014; Keidser *et al.*, 2015). and has been the target of research as a factor affecting performance in SiN for some time, since one of the subcomponents of the WM system; the phonological loop, is used for temporary storage of speech related information (Baddeley, 2003).

Age also affects WM performance (Gordon-Salant and Cole, 2016), but variation in WM capacity is seen in all age groups. Studies show higher scores in WM tasks predict higher scores in SiN tests (Van Engen, 2012). Gordon-Salant and Cole (2016) found SiN performance to be better in both young and old NH individuals with high WM capacity after controlling for age, however the relationship was stronger in the older age group.  A meta-analysis of studies looking at the effect of WM on SiN performance found no evidence of generalized WM as a reliable predictor of SiN performance in young NH individuals. There was however, a moderating effect of age (Füllgrabe and Rosen, 2016).

Other studies support the role of certain mediators of WM in improved performance in younger NH adults, such as cognitive inhibition (Stenbäck, Hällgren and Larsby, 2016). The role of WM may also be context specific, as it was elicited in studies using high context (SPIN) and predictability (HINT) speech material, but not with low context material requiring inference-making such as matrix sentences(Rönnberg, Holmer and Rudner, 2019). Hwang, Kim and Lee (2017) compared between NH and HI individuals from young, middle aged and elderly age ranges. They looked at their SiN performance, temporal resolution through a gaps-in-noise test, WM by a forward digit test and WM with attention by a backward-digits test. They found that a task incorporating WM with attention was the only predictor for variation in performance on a SiN test in the NH group, explaining only 13% of the variance in performance. In the HI group, increasing age was inversely correlated with performance on all tests and temporal resolution was the strongest predictor of performance on the SiN test explaining 33% of performance variance. This contrasts with Schoof & Rosen’s findings (2014). They did not elicit any difference in temporal processing between young and elderly NH adults.

Goossens *et al.* (2017) also compared between NH and HI individuals across age spans and found age to be the main predictor of performance decline in the NH group, especially in listening conditions with informational masking, even though subjectively the individuals weren’t complaining of any perceived difficulties. This decline may be explained by the variation in audiometric thresholds and the temporal processing deficits in the older age group. In this study, thresholds were measured up to 4kHz and thresholds up to 25 dB were accepted. Some individuals in the older NH group had thresholds up to 40 dB at 8kHz. The effect of age was not found in the HI group. Holmes and Griffiths (2019) found the highest predictor of variation in SiN performance in NH individuals to be age with a smaller association of cortical auditory grouping processes assessed by figure-ground perception tests. Selective attention also plays an important role in understanding speech in noise and is also subject to the effects of aging as it falls under executive functioning (Schoof and Rosen, 2014).

In the context of AFFD the aforementioned effects are important considerations due to the demographical changes in the modern industrialized era and the rise of an ageing workforce (Sluiter and Frings-Dresen, 2007).

### 2.4.7 Attention, motivation, and personality

Fluctuations in attention and motivation can affect psychophysical test performance (Fründ, Haenel and Wichmann, 2011). This can be related to fatigue, or altered medical, nutritional and sleep status (Brungart *et al.*, 2022). Laboratory simulations link sleep deprivation to decreased cognitive performance proportional to the amount of sleep deprivation. Fatigue adversely impacts learning and memory consolidation. In military environments, these factors may occur alone or with other stressors, such as fear or anxiety, and motivation can compensate in fatigued individuals only to a certain degree (Miller, Matsangas and Shattuck, 2008). Higher levels of anxiety can also impact attention and situational awareness affecting integration of sensory cues in extreme environments (McNeil and Morgan, 2010). In fast-response demanding jobs, required performance levels might need to account for these factors.

A growing body of literature supports the link between motivation and enhanced activity in attentional processing cortical regions during demanding tasks (Engelmann *et al.*, 2009). Personality trait conscientiousness in turn, is mediated by motivation in improved task performance (Ng, Ang and Chan, 2008; Richardson and Abraham, 2009). Other studies have found a direct associations between the salience/ventral attention network (SVAN) responsible for attention shifting and conscientiousness (Fleming, Heintzelman and Bartholow, 2016; Sassenberg *et al.*, 2022).

Personality theories describe the structures of enduring individual differences that distinguish people (Costa and Mccrae, 2012). These structures are stable over time, subject to effects of maturity with aging, an effect that predominates in the ages between 20-40 and gradually plateaus beyond that (Jackson *et al.*, 2009; Skoglund *et al.*, 2020). The development of personality theories surged 40 years ago, and several researchers independently came to realize that five factors were sufficient to measure individual differences (Costa and McCrae, 2012; Sackett *et al*., 2017). This became known as the five-factor model (FFM). The five-factor theory (FFT) is among the most famous empirical trait taxonomies of the FFM developed by McCrae & Costa (1985,1987,1996). The five dimensions of the FFT are Neuroticism, Extraversion, Openness to Experience, Agreeableness and Conscientiousness.

Personality assessments have a role in FFD in certain sectors such as law enforcement and medical practitioners with performance problems (Brown, Iannelli and Marganoff, 2017). Measures of personality have been recommended and are often included in neuropsychological assessments of service members in the US department of defence (DoD)(Kelly, Mulligan and Monohan, 2010) as mental health among military employee's was significantly associated with higher levels of conscientiousness (Lee, Sudom and Zamorski, 2013). Conscientiousness as described by Costa, McCrae and Dye (1991) has inhibitive and proactive aspects. Inhibitive aspects include moral scrupulousness and cautiousness. Proactive aspects include commitment to work and the need for achievement (DeYoung *et al*., 2007). Global trait conscientiousness is structured into six facets (Debusscher, Hofmans and De Fruyt, 2017):

- **Order:** the tendency towards organization and tidiness
- **Competence:** the sense of sensibility, capability, and accomplishment
- **Dutifulness:** adherence to social standards and norms of conduct
- **Achievement striving:** the level of striving and pursuing excellence
- **Deliberation:** the amounting of thinking, planning and caution before acting
- **Self-discipline:** the ability to persist in tasks despite boredom or distractions

Further research has focused on dividing each trait into two domains under which the facets are categorized. In conscientiousness it is industriousness and orderliness (DeYoung, Quilty and Peterson, 2007)(Roberts *et al.*,2005).

Trait conscientiousness may be of relevance to SiN test performance but has not been considered carefully. Since conscientiousness relates positively to the task performance aspect of job performance (detailed in the next section), it might be of importance to explore this in an AFFD context, because the way personality is expressed may change by virtue of training. This change over the course of training may obscure or exaggerate effects of auditory task performance in adverse conditions. The following section will discuss trait conscientiousness and its documented relationship with task performance and possibly AFFD.

## 2.5 Conscientiousness and task performance

Job performance can be divided into contextual, task and adaptive performance. Task performance encompasses behaviours and activities relating to core technical matters, contextual performance relates to behaviours that affect the social and organizational aspect of work and adaptive performance consists of the problem solving abilities that accompany ambiguous situations or changes in the work environment (Hackett, 2002). Conscientiousness relates positively to job (task and contextual) performance (Judge *et al.*,2008). It has also been shown to relate positively to higher academic performance and test scores ($r$= 0.29-0.39) across all educational levels and groups (Imhof and Spaeth-Hilbert, 2013; Stacey and Kurunathan, 2015). Regarding the conscientiousness -adaptive performance relationship, there are studies that found a predictive relationship for conscientiousness (Pulakos *et al.*, 2000) and others that concluded the predictive validity of conscientiousness did not extend to aspects with uncertainty requiring adaptability (Griffin and Hesketh, 2003; Thoresen *et al.*, 2004). Barrick and Mount carried out a meta-analysis of 117 studies obtaining sample-weighted mean correlations (corrected for unreliability) of .22 for Conscientiousness. This relationship was found across all occupational groups (Barrick & Mount, 1991). A correlation between conscientiousness and task performance has been demonstrated across studies ranging between 0.2-0.3 (Barrick 1991, Salgado 1997, Hurtz 2000), although the relationship is usually stronger between conscientiousness and contextual performance (Hassan, Akhtar and Yılmaz, 2016). This is understandable because task-performance is very job specific whereas contextual performance is universally similar (Bakker, Demerouti and Ten Brummelhuis, 2012). Jiang, Wang and Zhou (2009) examined the effect of conscientiousness on task and contextual performance in 478 participants. Regression analysis was conducted revealing that contentiousness was predictive of both task performance ($\beta = .14$, $p < .05$) and contextual performance ($\beta = .15$, $p < .05$).

Conscientiousness was positively related to performance on two out of three episodic prospective memory tasks (n= 141, OR= 2.22 and 1.92) (Cuttler and Graff, 2007). Graff and Cuttler postulated that the relationship may vary with task difficulty and be more prominent on more challenging tasks, an observation also made by Chen *et al*. (2001), while Barrick and Mount's findings related conscientiousness similarly to performance regardless of level of task complexity.

Research specific to the military sector has demonstrated a positive relationship between conscientiousness and military and academic training outcomes (Driskell *et al.*,1994; Bauer *et al.*, 2012; Bobdey *et al.*, 2021). Lin *et al.* (2019) found positive bivariate correlations (0.43, p<0.01) between conscientiousness and task engagement and performance only in high demand tasks involving military unmanned aerial system operations. The relationship did not extend to accuracy in task execution. Other military research has failed to predict competencies and skills from conscientiousness levels (Johansen, Laberg and Martinussen, 2014). According to the authors, given the importance of many conscientiousness facets in the military persona, they attributed the absence of the conscientiousness -performance relationship to the performance measures used in their study. Their study used self-reports, whereas many studies measure task-performance with more objective measures or external ratings.

The estimated validity scores of conscientiousness usually fall within a range of *r* = 0.2 and have been enthusiastically interpreted as predictors of job performance. Although they are small effects, they are consistent values arising across studies in the literature. The researcher of the current thesis acknowledges that it is a small effect with a moderate impact explaining a portion of the variation in job performance, but re-emphasize its stability and generalizability across occupations and performance measurement criteria (Hurtz and Donovan, 2000).

An ongoing debate is whether a stronger association exists between the global trait conscientiousness (Tett, Jackson and Rothstein, 1991; Stewart, 1999; Salgado, Moscoso and Berges, 2013; Debusscher, Hofmans and De Fruyt, 2017) or certain facets of the trait (Dudley *et al.*, 2006; Moldzio *et al.*, 2021). Harari, Naemi and Viswesvaran (2019) found global conscientiousness to be a more stable predictor across time, with facets more associated with momentary task performance, while Darr and Kelloway (2016) found facet achievement to be related to overall job performance (*r*=0.19, *k*=31) and task performance (*r*=0.13, *k*=26), with a stronger association to task performance than other facets.

It is undecided in the literature if the relationship with performance is general or context specific. Studies suggest the existence of context specific relationships, but most have looked at organizational culture contexts such as innovation-orientation (Wang *et al.*, 2012). From a military standpoint, Bilgiç and Sümer (2009) believe that context is vital in the personality-performance relationship and suggest studying the relationship in orientation with job-analyses. To further understand conscientiousness as a possible affecting factor on SiN/AFFD performance, assuming the possible existence of a context dependant relationship, there is an interest in exploring the relationship between conscientiousness and auditory test performance.

Many of the task-performance assessments are supervisory ratings of proficiency, competency, and requirement fulfilments, incorporating varying tasks across many occupations. The context of auditory tests is more basic and straightforward. It might seem questionable to extrapolate a connection between conscientiousness and auditory test performance from more generalized job tasks. However, there are studies of a more similar nature linking conscientiousness and sensory task-performance. A study by Zhang *et al.* (2017) on a military gunmen and civilian population showed the military group outperformed the civilian group on an optical illusion task (Muller-Lyer test) designed to assess visual misperception and scored lower in impulsive/sensation-seeking on a personality test, a trait encompassing facets of conscientiousness. This study, in addition to Cuttler and Graff's positive correlations between memory tests and conscientiousness as measured by the NEO-PI-R, mentioned earlier in this section allow the assumption of a possible relationship between auditory task performance and conscientiousness.

## 2.6 Choosing the right AFFD test

Fitness for duty tests fall under one of three categories, generic predictive tests (GPTs), task predictive tests (TPTs) or task simulation tests (TSTs) (Payne and Harvey, 2010).
**Generic predictive tests (GPTs):** tests with no job-related characteristics that are broadly applicable and may be predictors of FFD.

> *Example:* using a general SiN test, such as the matrix SiN test, to assess AFFD in the military.

**Task-related predictive tests (TPTs):** tests that may have job-related characteristics but are not based on a specific job task.

> *Example:* using the CRM SiN test in task-specific background noise to assess AFFD in the military. It has military command-type structure but is not task-specific.

**Task simulation tests (TSTs):** highly specific tests based directly on a job task.

> *Example:* using a virtual reality simulation of the air defence's command post environment with humming generator noise and listening to the exact commands given on the field then, based on the heard command, taking appropriate action with dummy missile ejector buttons.

An important consideration is the target group for which the standard is to be established. This decides the specificity of the standard. Using the military as an example, entry level recruits lack the experience and familiarity of seasoned employees rendering TSTs unsuitable. In their case, a more general standard such as a generic test format with military relevant elements would be more suited. In other cases, additional standards may be required supplemental to the general standards, e.g., certain military squads may be required to have higher levels of FFD (Robson *et al.*, 2017).

The characteristics of the occupational environment and occurring communications in addition to the interactive effects of experience and cognition are important considerations to allow judgement based on the holistic nature of the assessments as opposed to the isolated picture from a diagnostic test. All these factors interplay when deciding on a threshold in an appropriate test that separates unacceptable and acceptable performance levels (Casto and Cho, 2013).

Other important considerations are cost-effectiveness, practicality, and ease of test administration. These are areas where TST's, despite being the most sensitive and specific in terms of FFD assessment, fall short. Brungart *et al*. (2017) shortened the 200 SPRINT to 100 SPRINT but concluded that it probably wouldn't be used in the test battery being developed for the military as it has an open-set format and is susceptible to memorization and higher content learning. They concluded that a closed-format automated test would be preferrable for easier administration.

## 2.7 Development of relevant Arabic SiN test material

After reviewing available SiN tests (listed in Table 2.2) used or developed with the intent of use as occupational assessment tools, with respect to content and presentation, the Coordinate Response Measure (CRM) test, discussed in detail in section 3.3, was chosen as a guide to adapt an Arabic SiN test from. The reasons for choosing the CRM are as follows:

1. The CRM test is a communication performance task that measures speech intelligibility in a manner more congruent to receiving and comprehending orders, rendering it suitable for certain occupations requiring fast responses such as emergency services, military sectors, and industrial settings.
2. The material is easily adaptable. It is currently available in American and British English. The British English version was developed and validated by Semeraro *et al.* (2017).
3. Closed set word lists allow rapid easy data collection rendering it practical for screening purposes.
4. It can be adapted under different noise conditions.

The sentence structure of the British English CRM speech intelligibility test is "Ready call sign, go to colour number now". This differs from the original test first employed by Asher *et al.* (Moore, 1981) in which the call sign was fixed for each individual participant. The participants would then be tested simultaneously, requiring the participant hearing his or her call sign to respond. This presentation served to measure speech intelligibility in environments with multi-talker communications (Bolia *et al.*, 2000). The structure and presentation of the CRM fit the requirements of the intended Arabic SiN test in that the presented sentences resemble communications in military and emergency services.

## 2.8 Gap in knowledge

Currently no auditory fitness standards exist in Saudi Arabia other than PTA. Since it is not a suitable predictor of AFFD alone and given the paucity of language specific material for a more valid AFFD test, an Arabic SiN test will be developed then explored as a measure of AFFD.

The studies mentioned in section 2.5 documented a positive relationship between contentiousness and task performance in general and specific conditions. However, no studies to date explore the relationship between contentiousness and AFFD tasks. One research interest is to explore if conscientiousness influences SiN test performance in an auditory task.

To achieve this, speech material will be developed and implemented into a measure for AFFD. Performance on the developed measure and the effect of trait conscientiousness on performance will be studied.

## 2.9 Summary

This chapter gave an overview of hearing loss in the workplace with respect to its causes, effects, and assessments. Attention was directed to currents standards of AFFD and shortcomings, highlighting the situation in Saudi Arabia. More appropriate AFFD measures were considered and the rationale behind developing an Arabic AFFD measure was justified. Factors affecting SiN performance in an AFFD context were reviewed. Based on the available tests in the literature that have been used for AFFD and/or occupational testing purposes, the most suitable test; the CRM, was selected as a guide for development of the Arabic test.

Chapter three details the process of developing and optimizing Arabic speech material to ultimately be used in a SiN test.

# Chapter 3 Development of Arabic speech material for AFFD testing

## 3.1 Introduction

Functional assessment of SiN is an important component of AFFD testing. The lack of comprehensive AFFD test batteries, primarily the absence of Arabic test material and SiN testing in fast response occupations such as police force, military and emergency services is the driving force behind this research. The first phase of developing a measure of AFFD involves developing Arabic speech material to be used for SiN intelligibility assessments targeting command oriented, fast-response occupations.  As discussed in the previous chapter, the CRM was chosen to guide development of the speech material. This chapter will detail the development of the Arabic speech material corpus with the aim of ensuring equal intelligibility of all the words in noise prior to implementing them in an AFFD measure.

## 3.2 Aims and objectives

Aim one is to develop Arabic speech material for use in a measure of AFFD targeting certain populations, namely emergency services, and security forces such as law enforcement and the military. This will be achieved through the following objectives:

- Selecting speech material oriented to communications of the target populations.
- Recording the selected material and preparing it in a format suitable for administration as a test of intelligibility assessment.
- Assessing the intelligibility of the developed speech corpus by testing it on normal hearing subjects.
- Equalizing the intelligibility of the words based on the results of the assessment.
- Testing the equalized words to assess the effect of the adjustments made on the homogeneity of the word intelligibility.

Aim two is to ensure equalization of the words. This will be achieved by re-testing the words using a modified adaptive procedure.

## 3.3 Developing the test material

Developing speech material involves several stages. Decisions regarding material selection, recording, and presentation must be taken with careful consideration of the test purpose and nature of speech heard on the target jobs. This section will discuss the decisions taken and the rationales behind them.

### 3.3.1 Selecting the test material

Speech material selection depends on the test purpose and population. A sentence test was preferable to a word test, given its better simulation of real-world communication. Sentences also have more acoustic cues than words, making them easier for speech intelligibility in noise tasks (Sharma, Tripathy and Saxena, 2017). The CRM test is comprised of sentences with the following format: “Ready (call sign) go to (colour)(number) now”. In the original American English version, there are eight options for the call signs, four options for the colours and eight options for the numbers, one through eight. This results in a total of 256 possible sentences. Originally it was recorded by eight different talkers resulting in a total of 2048 sentences. It has been used in studies on informational masking and is advantageous in its limited context and closed format (Eddins and Liu, 2012). The general structure of the CRM was chosen as a guide to test development due to its close representation of communicating orders, which would be required in all targeted occupational settings. All resulting sentences were syntactically but not necessarily semantically correct.

Having chosen the CRM as a starting point for developing the Arabic sentence structure, the speech material was oriented with likely target groups such as the military in mind. Word groups and sentence structure were selected based on the feedback of subject matter experts (SMEs) in the Saudi Arabian military. Information was gathered from three SMEs over several telephone communications and emails. The words chosen were simple and common, fitting the purpose of speech intelligibility testing.

The structure of the British English CRM test is “Ready call sign, go to (colour) (number) now”. Call signs in the Saudi Arabian military are in the form of Arabic letters or names. The Arabic alphabet consists of 28 letters, 27 of which are monosyllabic. From these 27 letters, the 15 most frequent letters were chosen and divided into 2 groups: CVC letters (nine) and CV letters (six). The rationale for dividing them was that the other word lists ranged from six (the digits list) to nine items (the directions list) and having all 15 letters would have created a large discrepancy between the lists. Also, phonemic differences exist between the CV and CVC lists. Based on mapping letter occurrence frequency from the full text of the Quran, all the high frequency occurring letters were included, except the most frequent letter of the Arabic alphabet (*alif*) as it is disyllabic (Brierley *et al.*, 2016).

From the digits zero to ten, disyllabic numbers were chosen as they constituted the majority. Colours are not used in Saudi military communications and were substituted with the disyllabic directions and commands listed in Table 3.1.

***Table 3.1*** *Word lists initially chosen for Arabic speech test material*

| Letters | | Digits | Directions and commands |
|---|---|---|---|
| CVC | CV | | |
| جيم Djeem | باء Baa | واحد Wahid (one) | وراء Waraa (behind) |
| لام Laam | تاء Taa | اثنان Ithnan (two) | أمام Amam (in front) |
| ميم Meem | راء Raa | خمسة Khamsa (five) | يسار Yasar (left) |
| نون Noon | ياء Yaa | ستة Sitta (six) | يمين Yameen (right) |
| واو Wow | حاء Haa | سبعة Sabaa (seven) | خطوة Khutwa (one step) |
| كاف Kaaf | خاء<br>Khaa | تسعة<br>Tisaa (nine) | شمال Shamal (north) |
| دال Daal | | | جنوب Djanoob (south) |
| سين Seen | | | سريع Saree (fast) |
| شين Sheen | | | بطيء Batee (slow) |

A closed set matrix format was chosen for the presentation of the word lists. Closed sets restrict the test taker to choosing an answer from a limited known number of options as opposed to an open set format where the test set is unknown. While the closed-set format may have a higher guess rate, it is easier to administer and less time consuming. The matrixed format of the test also renders the sentences more unpredictable than previously predefined sentences (Kollmeier *et al.*, 2015). Due to the scarcity of Arabic SiN material, development of an easy to administer test that does not require a specialized administrator is logical, to gain maximum future benefit of the test.

### 3.3.2 Recording the test material

The purpose of recording was to obtain speech material that would ultimately be used in a measure of AFFD for a specific target population. The populations in mind were those in command oriented fast-response occupations such as the military and police force. Many of the communications in their jobs are in the form of instructions vocalized with urgency and command. Thus, it was required that the resulting recordings have the same qualities. Recordings of two male speakers were made. Since the recording venue and equipment were made available for a limited time, the decision was made to record two speakers to have more than one selection in case one of the speakers' recordings did not meet all requirements. Male voices were chosen because certain sectors in Saudi Arabia such as the military, emergency services, factories and civil aviation sectors are male dominated with some sectors having only male personnel.

Recordings were done during September 2018 in a small anechoic chamber (room 1051, building 15 on the Highfield campus of the University of Southampton). The chamber's dimensions are 5.2 x 5.1 x 2.8 metres, and the walls and ceiling were lined with open cell polyurethane foam wedges, 30cm long with a 30cm square base. The floor was also covered with foam wedges to minimize reverberation in the chamber. The inner walls were isolated by a 25mm air gap from the outer structural wall (Semeraro, 2015). The speakers were sat in the chamber, as shown in Figure 3.1, 50cm away from two microphones, a Bruel and Kjaer type 4189-L001 and a Rode M5. The justification for using two microphones is that they both had distinctive audio features which made it very difficult to choose one prior to recording. The Rode M5 is frequently used for speech recording purposes, but the Bruel and Kjaer is known to have sharper overall sound quality. Both microphones were connected to a Babyface RME soundcard. The soundcard was connected to a MacBook Air (1.7 GHz Intel Core i5 processor, OS X 10.9.5 software). Recording was done using Adobe Audition (version 5.0.2 Build 5). Sampling was done at a rate of 48,000 Hz.

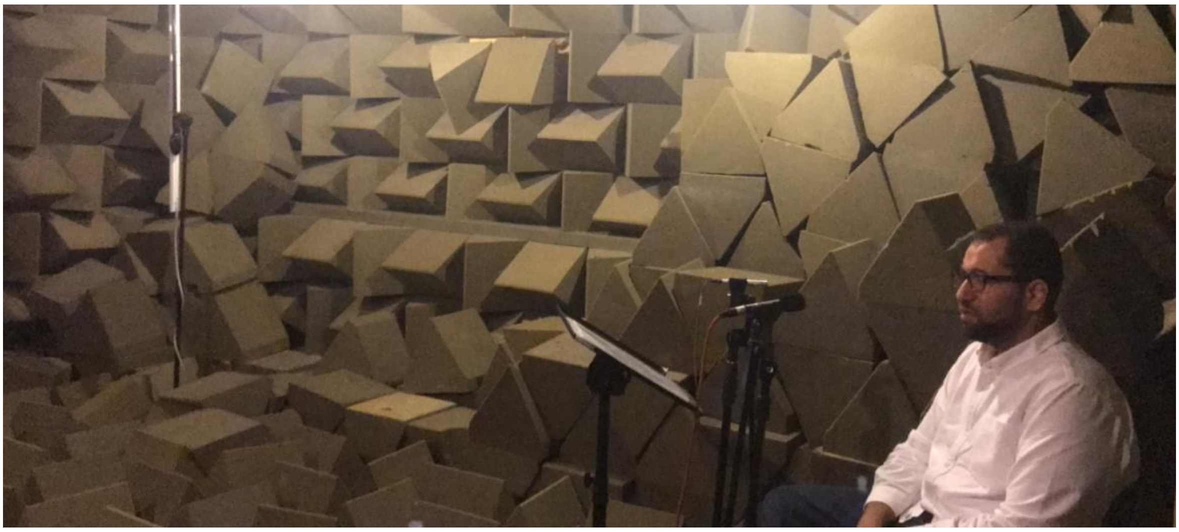


***Figure 3.1*** *Setup for recording the Arabic speech corpus in the anechoic chamber*

The speakers were given three wordlists consisting of letters, digits, and directions. They were made aware of the purpose of recording, and that the material needed to be vocalized in the same tone with a steady voice quality. Given the instructive nature of the test sentences, it was required the material be delivered with firmness. The speakers were instructed on the desired pronunciation and intonation. They performed two test runs before starting the actual recordings.

They were encouraged to maintain a constant tone and vocal effort. Subsequently, recordings of the Rode M5 microphone were chosen as they had better speech sound quality, and first speaker recordings were chosen for voice clarity and intonation better matching the purposes of the test.

Each wordlist was done as one recording with one second duration pauses between the words. This was preferred to individual recordings of each sentence as it allows testing the intelligibility of the words individually in noise and eliminates effects of coarticulation and prosodic cues. Coarticulation is the effect a word has on the pronunciation of the word preceding or following it. Since the nature of orders is slightly different than that of normal conversational speech, the coarticulation effect was not desired.  Recordings were later chopped into individual words using Adobe Audition and further adjusted using MATLAB to have the same Root Mean Square (RMS) amplitude.

### 3.3.3 Choosing a suitable psychophysical testing method

The method chosen to initially test and equalize the speech material is the method of constant stimuli (MoCS). The MoCS uses the same set of stimuli (usually consisting of five to nine values) repeatedly throughout the test. The lower end represents a stimulus that can almost never be detected, and the upper end represents a stimulus that is almost always detected. The threshold is located within the range of stimulus values (Gescheider, 1997). This method is conventionally used in test development phases as the distribution of the gathered data points is sufficient and spread out enough to plot psychometric functions (Leek, 2001). Its disadvantages are that it requires numerous trials to gather data (Gelfand, 2004) and is subject to floor and ceiling effects. The ceiling effect is a measurement limitation that is observed when the participant's score is too high due to the upper limit being too low. The floor effect is seen when a tests lower limit (floor) is too high resulting in the participants scoring near the bottom (Taylor, 2010).

Since precision in the method of constant stimuli increases with increased trials, the trials should be no less than approximately 410 points (Kingdom and Prins, 2010). The levels tested are represented as distributed points on a function. The collected data of this psychophysical method are expressed as a psychometric function (PF) (Kingdom and Prins, 2010).

A PF describes an individual's performance on a psychophysical task, e.g., responding every time they hear a word, by plotting the responses to the task at different stimulus intensities. Through quantitative measurement of the relationship between a physical stimulus and the perceptual response, stimulus levels required to achieve a certain measure of performance can be deduced (Wichmann and Hill, 2001). The important points of a PF are illustrated in Figure 3.2. Fitting a PF requires adjusting the parameters through careful selection of the stimulus levels to minimise errors in the data, as one of the common causes of poor fit of data is poorly chosen stimulus levels. This emphasizes the importance of pilot studies in ensuring selection of adequate levels.

Four important parameters describe the PF, two changeable parameters related to the underlying sensory mechanism (location α and slope β), and two fixed parameters required to complete the non-sensory description of the function by defining the upper and lower limits of the function (**λ** and **γ**).

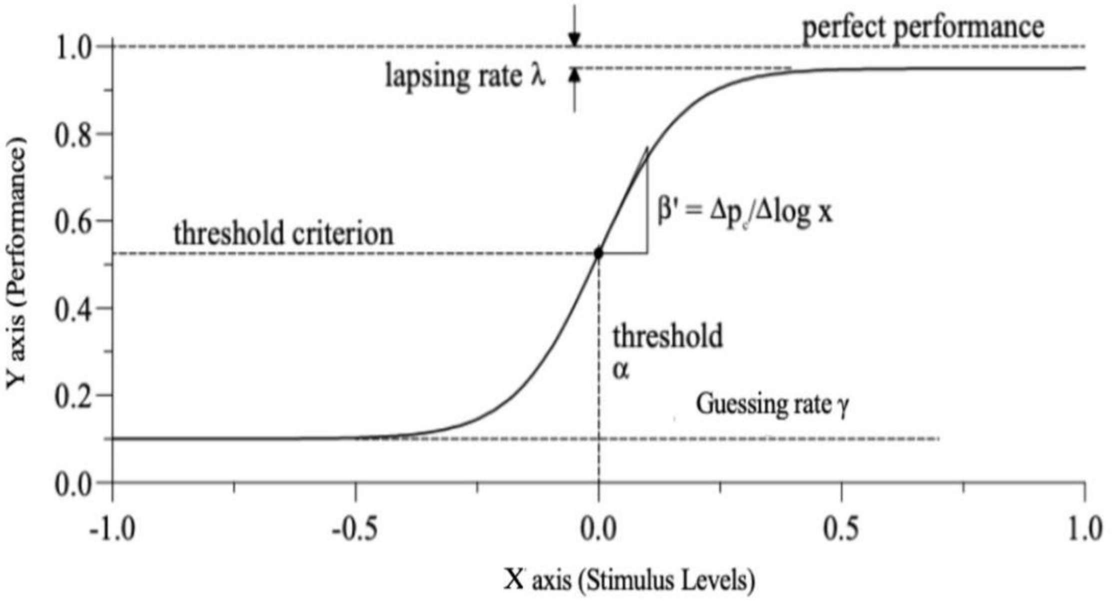


***Figure 3.2*** *Diagram of the representative points of a psychometric function, adapted from (Strasburger, 2001)*

α is the position along the abscissa and is defined as a certain value, in our study it corresponds to the inflection point which is termed the 'location'. The location is the steepest point in the slope, and it is an independent variable. Although many researchers express the level of performance using the speech recognition threshold (SRT) at a certain percent, this is not favourable as the SRT is a function of all the PF parameters. For this study, we are interested in the sensory mechanism underlying the performance which is described in terms of location α and slope β.

β corresponds to the slope. The slope expresses the ratio of maximum change in performance (Ay) to the change in stimulus level intensity (Ax). It is a measure of the function's precision (Kingdom and Prins, 2010) and is expressed as some form of Ay/Ax (%/dB). (Wilson and Carter, 2001). Steep slopes indicate large changes in intelligibility occurring due to small changes in SNR while shallow slopes indicate that larger changes in SNR are required to elicit the same change in intelligibility as the steep slope (Macpherson and Akeroyd, 2014). Slopes can be used as comparators between different models provided conversions between slope functions are applied.

***γ*** is the guess rate. It corresponds to 1/*n* (with *n* being equal to the number of options per trial).

***λ*** is the lapse rate. It corresponds to the deviation from a perfect score due to a lapse of attention, judgement, or another cause. Fixing the lapse rate at a very small value, for our study we have chosen to fix it at 0.01, will prevent significant effect of bias on the location and slope.

### 3.3.4 Participants

Participants included in the study were otologically healthy, native Arabic speaking adult civilians between the ages of 18-45 years. Volunteers were screened for otological or chronic medical illnesses by a questionnaire. Individuals with chronic medical illnesses such as diabetes or autoimmune diseases were excluded to ensure the absence of disease-related neuropathies that could affect hearing (Baiduc and Helzner, 2019; Jeong *et al,* 2019). The inclusion criteria were stringent to eliminate variability between subjects as much as possible and ensure audiological fitness for the test development process.

The aim at this stage was to gather data sufficient to estimate the PFs of the words and equalize their intelligibility to ensure a homogenous speech corpus. The number of required participants was based upon similar studies developing speech material, as predefined statistical sample size calculation techniques currently do not exist for this context. Sample sizes in previous studies ranged from 20 to 50 participants (Smits *et al.*, 2004; Ozimek *et al.*, 2009; Brungart, Sheffield and Kubli, 2014; Houben *et al.*, 2014; Vaez, Desgualdo-Pereira and Paglialonga, 2014; Semeraro *et al.*, 2017). A total of 20-25 participants were to be recruited for the first study of this experiment. Recruitment was done using emails advertising the study. Participants were King Abdul-Aziz university students/staff and people from outside the university. The participants were different in each session except for three individuals, who took part in sessions two and three. This allowed the sample to be more representative of the variation in the population.

Prior to the first session, participants were screened by the experimenter using an otological health questionnaire and PTA following the standard clinical procedure for PTA in accordance with the British Society of Audiology (BSA) recommended procedures (2018). Participants passing the screening (responding reliably at ≤15 dB HL from 0.25-8 kHz) in both ears proceeded to the speech task. Since the upper limit of normal hearing thresholds is 20 dB, a threshold limit of 15 dB was chosen to ensure none of the participants had borderline thresholds.

## 3.4 Overview of the Test Procedure

The experiment, approved by both the University of Southampton (ERGO ref: 45925, Appendix C) and King Abdul-Aziz University (Appendix D), consisted of three sessions. Each session lasted approximately one hour. Participants were given a break every 20 minutes during testing, or more frequently if required to avoid fatigue or boredom.

All sessions had identical structure. Participants were sat at a table with the word lists in front of them. The test was administered binaurally through circumaural headphones. Binaural presentation was preferred to monaural as it bears more resemblance to real world listening and results in better performance (Theunissen, Swanepoel and Hanekom, 2009). Headphones were preferred to loudspeakers as they result in less variability. They listened to the recorded words preceded by a carrier phrase “ready”. For the background noise, using a specific MATLAB code, SSSN was generated by digitally filtering gaussian noise to have the same LTASS of the recorded words, concatenated together. Stationary noise has less fluctuations in the noise level, giving more comparable results, making it better suited for this developmental phase (Hagerman, 1997; Killion *et al.*, 2004; Wagener and Brand, 2005b).

This word presentation and background noise selected was strictly for the purpose of equalizing the intelligibility of each word before presenting the words in a sentence format. The lists were presented at a SNR higher than the selected SNR’s for the test (0 dB (A)) first for familiarization. The test lists were then presented in random order. The random list order was generated through a website (Haahr, 2019) using a model that creates true randomization based on a naturally occurring phenomenon (in this case, atmospheric noise). The participant was required to respond with the answer they heard from the available options and to guess if they were unsure. The researcher recorded the answers on the computer.

The noise exposure levels did not exceed the sound level which defines an unusual experiment as outlined in the Institution of Sound and Vibration Research (ISVR) Report 808- info for noise and vibration ethics. The noise exposure calculation is based on both the target sentence and the masker never exceeding 70 dB (A) independently, and so when both the target and masker are presented together the highest maximum combined level will be approximately 73 dB (A). The exposure duration is based on the participant having a maximum listening time of 2 hours in any 24-hour period.

Calibration of the stimulus was measured through the portable appliance testing (PAT) tested circumaural headphones used for testing (Sennheiser HD650). The calibration stimulus is speech shaped noise with the same frequency shaping as the test sentence stimuli. The noise is presented at the peak level the sentence stimuli will reach, ensuring that the maximum presentation level is checked. Headphone output was measured using an ear simulator and through a calibrated sound level meter. Objective calibration was done weekly. Subjective listening checks were carried done prior to each test session to check levels and sound quality.

## 3.5 General strategy for analysis

The collected data are expressed as a PF. In this section, the strategy used to obtain the PFs will be explained in addition to the important elements for analysis. Using a code written by Dr Daniel Rowan run in MATLAB v2018 software, data were collected from the participants as trials correct for each word per SNR. In all test sessions, the number of presented trials was never less than 800 trials. The following steps were followed to express the functions.

1. Raw data was prepared, and a suitable function was fit.
2. The functions goodness of fit was determined
3. Accuracy of the estimates of parameters was determined
4. Similarity of thresholds was evaluated
5. Steepness of slopes was evaluated
6. Level adjustments were calculated as required

steps 1,2, and 3 used the Palamedes toolbox in MATLAB v2018. Each step will be explained below.

### 3.5.1 Preparing the raw data and fitting the function

Data from all participants were pooled and the average proportion correct per SNR was calculated from the pooled data. From these averages, a logistic function was fitted by the method of maximum likelihood. Logistic functions are convenient and commonly used for speech intelligibility assessment purposes (Pichora-Fuller, Schneider and Daneman, 1995; Treutwein and Strasburger, 1999; Wichmann and Hill, 2001; Macpherson and Akeroyd, 2014; Semeraro *et al.*, 2017).

### 3.5.2 Determining Goodness of fit of the function

Goodness of fit was assessed by doing a parametric bootstrap analysis. Bootstrapping is a statistical method that resamples the existing data multiple times to generate empirical estimates of the sample distribution (Field, 2009). This was expressed by the p-value for the measure of deviance (pDev), a value between 0-1. This measure depends on the distribution of the data points and shows us how well the PF captures this data. It is conventionally agreed upon in the literature that the higher the value the better the fit of the data, with values below 0.05 expressing unacceptably poor data fits (Kingdom and Prins, 2010). A p-value > 0.8 was accepted as a very good fit.

### 3.5.3 Determining accuracy of estimates of parameters of fit for the function

Standard errors (SE) and their confidence intervals (CI) were considered alongside the pDev for accurate interpretation of the results. These values are also derived from the bootstrap analysis. The SE measures the distance of the data points from the function, capturing the precision of the models prediction (Frost, 2019). The smaller the distance of the data from the PF i.e., the smaller the SE, the closer the data points are to the function. Standard errors are influenced heavily by sample size and the amount of variation in the data. Acceptable SEs were < 3dB SNR for the location and <5 for the slope. The 95% CI for the resulting locations is an interval generated by a procedure which in repeated sampling has at least a 95% probability of containing the location's true value, for all possible values (Neyman, 1937). It depends on the distribution of stimulus values x. Contrary to common belief, wider intervals do not always indicate less precision in the results (Morey *et al.*, 2016).

### 3.5.4 Checking similarity of the word 'locations'

The threshold or 'location' in the context of our study refers to the inflection point, which is the steepest point in the slope. Many speech test development studies allow a maximum range of adjustment (usually 3-4 dB), and while this arbitrary range is conventionally accepted, Kollmeier *et al.*(2015) pointed out that language differences must be taken into consideration when determining the acceptable ranges . Since studies developing Arabic speech test material are scarce, there was no reference for the acceptable range for adjustment specific to the language, so an adjustment of up to 4.5 dB was accepted.

### 3.5.5 Checking steepness and similarity of the 'slopes'

The higher the slope value, the steeper the slope. Slope values <0.37 (corresponding to 9%/dB) were considered shallow. Values >2 (50%/dB) were considered an improper fit. The MATLAB code used for analysis, created based on a method described by (Kingdom and Prins, 2010) generates a slope value without any units. To express the value as %/dB the following equation was used to calculate the slope value (Strasburger, 2001):

$$b' = ((1 - g/4).b$$

$$b' = ((1 - 0.01/4).b$$

$$b' = (1/4).b$$

$$\%/dB = b' * 100$$

Where b’ is a numerical value derived from the slope and independent from threshold and g is the guess rate, which in this study is a value equal to 0.01.

### 3.5.6 Calculating level adjustments, if required

To equalize the intelligibility of test material, a set criterion must be chosen, and corrections must be applied to the material to equate it to the set criterion. In this study, mean location was chosen as the criterion. Equalization was done by adjusting the RMS amplitudes of the words to match the mean location. This was calculated from the difference between the location of the word and the average location for the word list. Based on the calculated difference, amplitudes of the words with a location lower than the average location were reduced and words with higher locations were increased (Warzybok, Zokoll and Kollmeier, 2016). The adjustments were limited to a maximum of ±4.5 dB SNR.

## 3.6 Study one: Equalizing the intelligibility of the speech material

This study consisted of three sessions. Session (1.1) assessed the intelligibility of the words. Based on the results, the word choices were revisited and modified. Session (1.2) was conducted to assess the intelligibility of the words after making necessary changes to the speech corpus. Based on the results, level RMS amplitude adjustments were applied, and session (1.3) was conducted to assess the intelligibility of the words after RMS equalization. The sessions are summarized in Figure 3.3. In the following sections, the method and results of each test session are detailed.

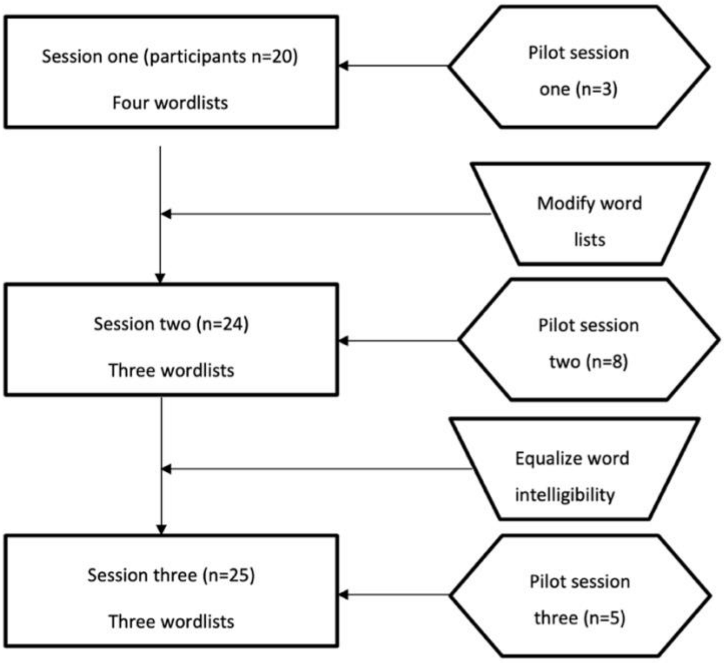


***Figure 3.3*** *Flowchart summarizing the stages of study one*

### 3.6.1 Study one-session one (1.1)

The aim of this session was to assess the intelligibility of the recorded words to calculate the adjustments required to equalize the intelligibility of the words from the resulting PF's.

A pilot was done to investigate whether the chosen stimulus levels were sufficient to generate a complete PF. Based on a similar material development study (Semeraro, 2015), the following SNR levels were chosen: -6, -10, -12, -14, -16, and -18 dB SNR.

Three participants were tested and based on the results, the following SNRs were chosen:

- The 'CVC' letters (list one) were tested at -8, -10, -12, -14, -16- and -18 dB SNR
- The 'CV' letters (list two) were tested at -4, -8, -10, -12, -14- and -16 dB SNR
- The 'digits' (list three) were tested at -10, -12, -14, -16, -18- and -20 dB SNR
- The 'directions' (list four) were tested at -8, -10, -12, -14, -16- and -18 dB SNR

**Results for study 1.1**

Data collection began at the end of December 2018. Data were gathered from 20 participants in a quiet booth located in the audiology clinics at King Abdul-Aziz University in Jeddah. Using the code in MATLAB, the following results were obtained: Mean locations, listed in Table 3.2, varied considerably between lists, ranging from -8.66 dB SNR for the 'CV letters' to -17.22 dB SNR for the 'digits; For the 'CVC letters', locations were within a range of ±8.06 dB SNR with the word 'seen' deviating from the mean by -8.9 dB SNR and ‘kaaf’ differing from the mean by 7.16 dB. SNR. Slopes were steep apart from ‘laam’ (7.78%/dB) and ‘wow’ (8.03%/dB).

For the ‘CV letters’, the mean location was -8.66 dB SNR, higher than the other word lists. The word ‘baa’ differed from the mean location by 4.72 dB SNR. All slopes were shallow except for ‘khaa’ (10.21%/dB) and ‘yaa’ (11.92%/dB). Mean location for the ‘digits’ was -17.22 dB SNR, lower than the other lists with steep slopes. The locations for the ‘directions’ ranged within ±8.08 dB SNR with ‘shamal’ deviating from the mean location by 11.98 dB SNR. It was also the only word exhibiting a shallow slope (3.41%/dB).

***Table 3.2*** *Mean location for each wordlist in study 1.1*

| Average location for each word list-session one dB (SNR) | | | |
|---|---|---|---|
| **CVC letters** | **CV letters** | **Numbers** | **Directions** |
| -13.13 | -8.66 | -17.22 | -13.10 |

The boxplots in Figure 3.4 show the relationship between the distribution of the locations of the words in each list. A ceiling effect can be for the word ‘seen’ in list one, and a floor effect can be seen for ‘shamal’ in list four.  There is a wide range of variability in the intelligibility of the lists. Given the locations of several words across the lists differed from their means by values exceeding the accepted range within which level adjustments could be applied (±4.5 dB SNR), further analysis was not done, and the word choices were revisited.

**Modification of the word lists**

The ‘CV letters’ were removed for being too difficult. Difficulty can be attributed to the nature of the words all ending with the same vowel and differing only in the consonant. The ‘digits’ were removed for being too easy, with the average location for the words ranging from -16.5 to -19.7 dB SNR. In the ‘CVC letters’, the letters *daal, seen and sheen* were removed and replaced with *zen, ein and saad*. In the ‘directions’, the words *shamal, waraa, and yasar* were removed and replaced with the words *baeed, gareeb and aali.*

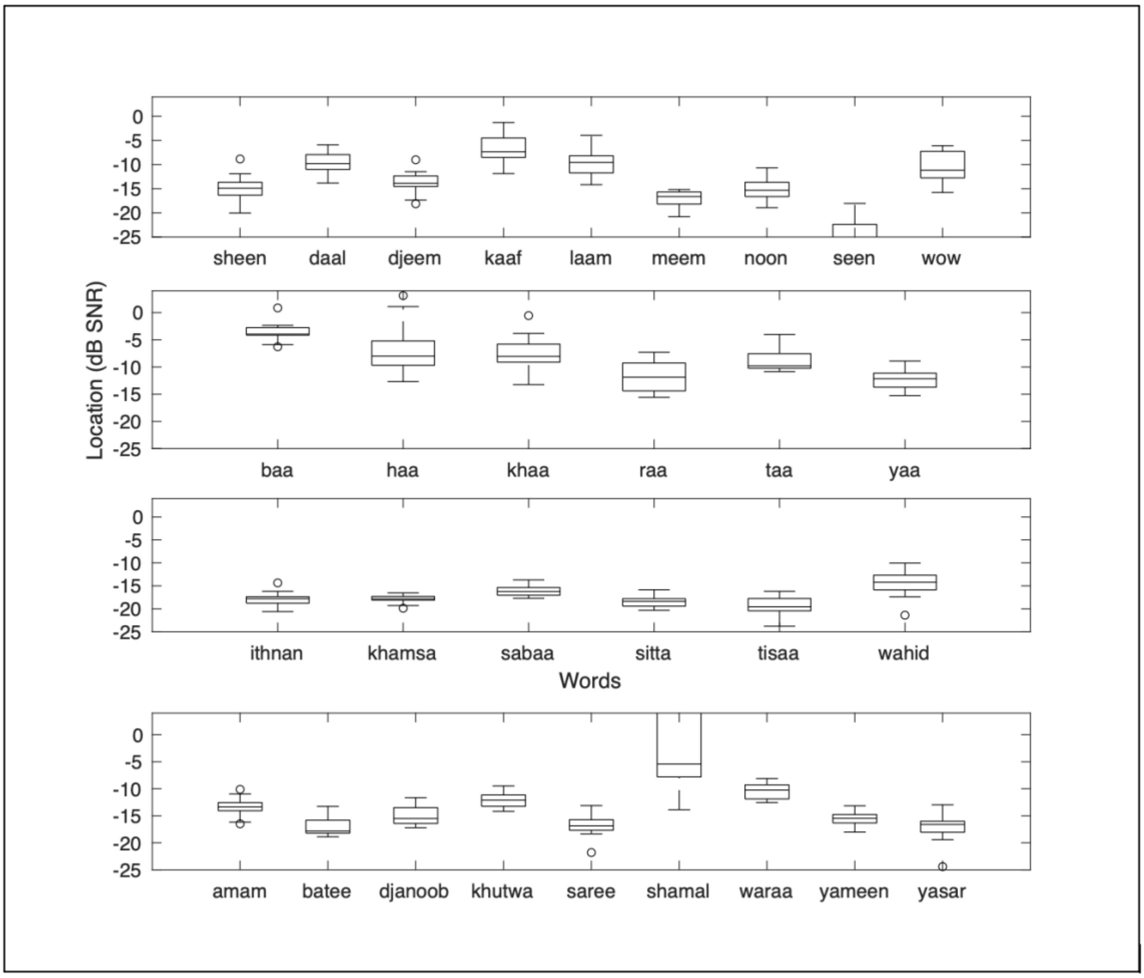


***Figure 3.4*** *Boxplots of word locations in (dB SNR) for each word list in study 1.1*

As two lists would be insufficient for test material, a third list was required. The second option for call signs; codenames, was chosen. A list of commonly used codenames was compiled, and after selecting the suitable options based on the tests criteria (same number of syllables), a list with 13 codenames was generated. The words chosen and recorded for session two are listed in Table 3.3. The new words were recorded by the same speaker of the previous recordings in a sound treated booth with the same equipment and conditions used for obtaining the previous recordings, discussed in section 3.3.2. The recordings were chopped, amplitude adjusted and embedded in MATLAB noise following the same procedure used for the previous recordings.

***Table 3.3*** *Modified wordlists for study 1.2*

| letters | Codenames | Directions and commands |
|---|---|---|
| جيم Djeem | أسد Asad (lion) | أمام Amam |
| كاف Kaaf | فرس Faras (mare) | بطيء Batee |

| letters | Codenames | Directions and commands |
|---|---|---|
| لام Laam | نمر Nimir (tiger) | جنوب Djanoob |
| ميم Meem | فهد Fahad (panther) | خطوة Khutwa |
| نون Noon | مطر Matar (rain) | سريع Saree |
| زين Zen | جبل Jabal (mountain) | قريب Gareeb |
| عين ein | هيثم Haytham | بعيد Baeed |
| صاد Saad | صقر Sager (falcon) | عالي Aali |
| | سهم saham (arrow) | |
| واو wow | رعد Raad (thunder) | يمين yameen |
| | بدر bader (moon) | |

### 3.6.2 Study one-session two (1.2)

This session aimed to assess the intelligibility of the modified word lists in reference to the same set criterion previously used (mean location) to make the necessary changes for intelligibility equalization. To equalize the words, RMS amplitudes were adjusted to the average location $\alpha$ of each list, taking into consideration adjustment of the average across lists.

Piloting was done to select the suitable SNRs for each list and exclude words with undesirable response patterns not fitting with the rest of the responses. The pilot was done on eight participants at King Abdul-Aziz University in April 2019 in a quiet room in the Audiology department. Based on the results, all the words in table 3.3 were included and the following SNRs were chosen:

- The ‘letters’ (list one) were to be tested at -3, -8, -12, -14, -16, -18 dB SNR
- The ‘codenames’ (list two) were to be tested at -5, -10, -12, -14, -16, -18 dB SNR
- The ‘directions’ (list three) were to be tested at -6, -10, -12, -14, -16, -18 dB SNR

**Results of study 1.2**

Twenty-four new participants took part in session two. They were recruited and screened using the same method explained in section 3.3.4. Testing was done in a quiet room in the Audiology department of King Abdul-Aziz University Hospital in Jeddah. The test was conducted throughout May 2019. Participants underwent 840 trials in this session. The results are reported below.

### Locations

The ‘letters’ exhibited the highest variation in intelligibility, with a mean location of -12.48 dB SNR. This variation can be seen in the PFs in Figure 3.5. The words ranged from the highest location for ‘wow’ at -8.06 dB SNR to the lowest location for ‘zen’ at -15.39 dB SNR. The ‘codenames’ ranged within ±2.5 dB SNR with the highest location -15.62 dB SNR for ‘asad’ and the lowest -10.54 dB SNR for ‘raad’. The mean location for ‘directions’ was -12.99 dB SNR. Locations ranged within ±4.22 dB SNR, with ‘gareeb’ having the highest location at -9.24 dB SNR and ‘saree’ having the lowest location at -17.68 dB SNR.

### Slopes

Slopes for the ‘letters’ were all steep except for ‘wow’ and ‘kaaf’ (58.5%/dB and 50.5%/dB respectively) indicating improper fit. The ‘codenames’ exhibited steep slopes except for ‘saham’ which visualized in Figure 3.6, had a shallower slope of 8%/dB. All slopes for ‘directions’ were steep.

### p-value for the measure of deviance and standard errors

All words exhibited a very good fit except for ‘khutwa’ and ‘saree’ from the directions list (pDev = 0.76) which is still acceptable. Standard errors for locations of the lists were all <2 dB SNR, apart from ‘wow’, ‘saham’ and ‘saree’. All slope SEs were <5 except for ‘gareeb’.

### Comparison between the lists

The mean locations for the lists were all within 1dB range indicating similar levels of difficulty across the lists. All words were equalized in intelligibility by adjusting the RMS amplitudes as previously explained in section 3.5.

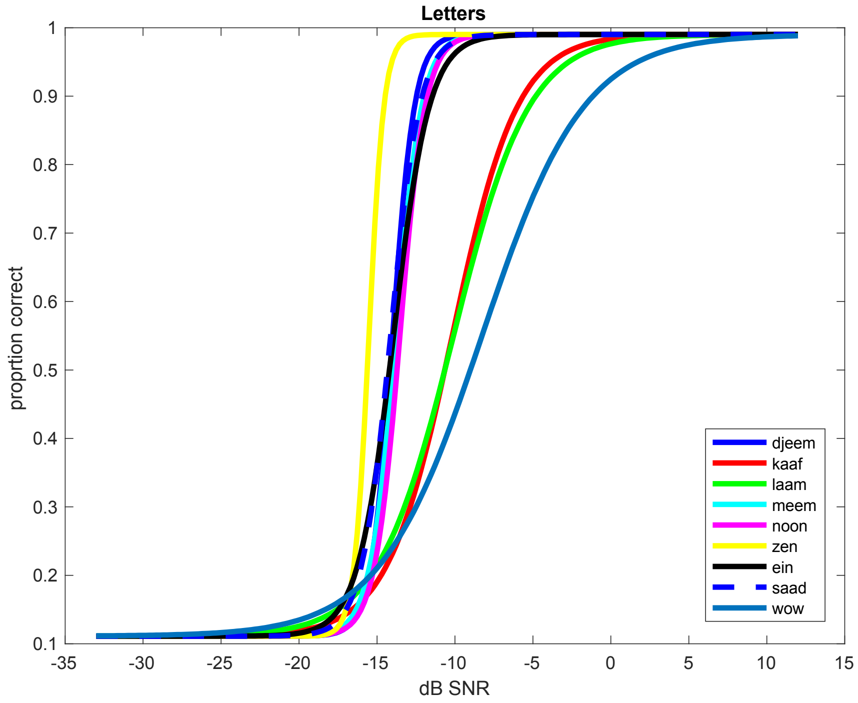


***Figure 3.5** Logistic functions for the 'letters' in study 1.2 (obtained from the mean threshold at each SNR)*

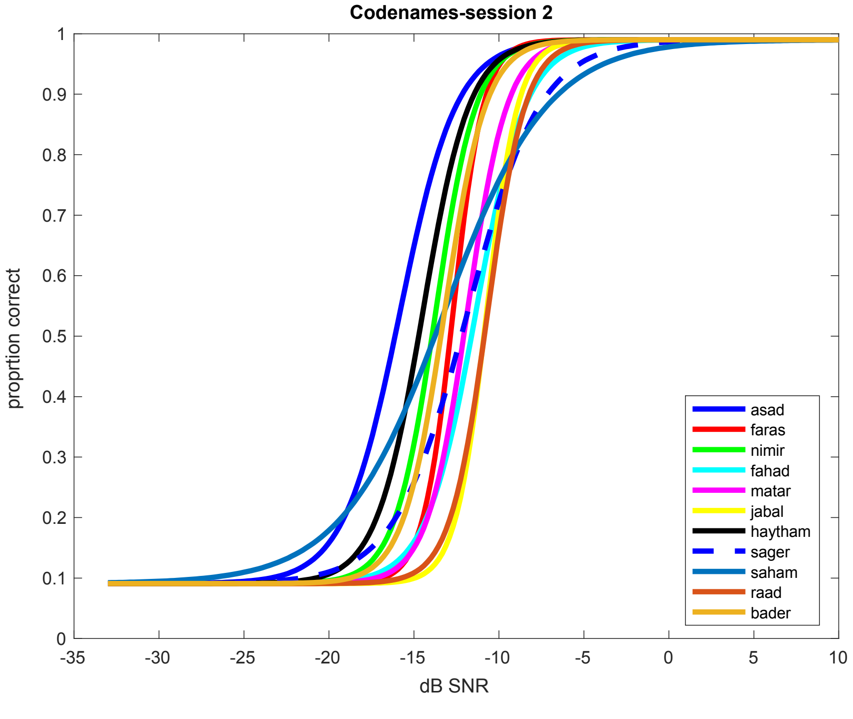


***Figure 3.6** Logistic functions for the 'codenames' in study 1.2 (obtained from the mean threshold at each SNR)*

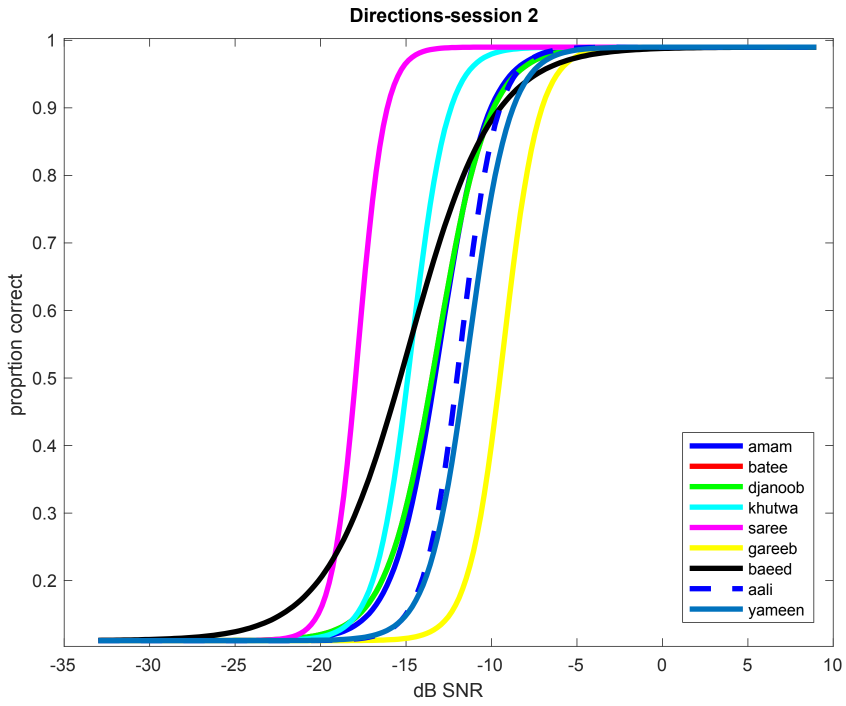


***Figure 3.7*** *Logistic functions for 'directions' in study 1.2 (obtained from the mean threshold at each SNR)*

### 3.6.3 Study one-session three (1.3)

The purpose of this session was to determine the homogeneity of the words after equalizing intelligibility. Piloting was done to ensure none of the words sounded unnaturally loud or quiet after equalization compared to the remaining words. Also, based on participants' results and feedback regarding the difficulty of the words, two words were removed from the 'codenames' list to homogenize the number of words across all lists.

The pilot was done in a quiet room at the University of Southampton in the beginning of July 2019. The equalized words were piloted on five native Arabic Saudi participants. Based on their results, the following decisions were made: The words 'Asad' and 'Haytham' were removed from the 'codenames' list for being too easily recognized, resulting in a list consisting of nine words.

Level adjustments were applied to the remaining words within and across lists. Table 3.4 shows the final level adjustments for equalization. Level adjustments preceded by the (-) denoted that the RMS amplitude of the word was decreased by this amount. Level adjustments not preceded by the minus sign denote that the words amplitude was increased by this amount. Testing was conducted at the following SNRs for all lists: -4, -8, -10, -12, -14- and -16-dB SNR.

***Table 3.4*** *Level adjustments required for equalization of words to be tested in study 1.3*

| Applied level adjustments in dB | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Letters** | *djeem* | *kaaf* | *laam* | *meem* | *noon* | *zen* | *ein* | *saad* | *wow* |
| | -1.4 | 2.4 | 2.5 | -1.2 | -0.9 | -2.9 | -1.3 | -1.5 | 4.3 |
| **Names** | *faras* | *nimir* | *fahad* | *matar* | *jabal* | *sager* | *saham* | *raad* | *bader* |
| | -0.1 | -1.1 | 1.3 | 0.7 | 1.9 | 0.8 | -0.6 | 1.9 | -0.6 |
| **Directions** | *amam* | *batee* | *djanoob* | *khutwa* | *saree* | *gareeb* | *baeed* | *aali* | *yameen* |
| | -0.4 | -0.5 | 0.7 | -2.1 | -5.2 | 3.3 | -2.2 | 0.8 | 1.3 |

**Study 1.3 results**

Twenty-five participants took part in this session, three of whom had participated in session two. Each participant underwent 810 trials during the test session, completing the test within an hour excluding breaks. Testing was done using the same venue and protocol of the previous sessions.

**Locations and slopes**

The ‘letters’ still exhibited some variability as seen in Figure 3.8, with a mean location of -13.42 dB SNR.  The words ranged within ±3.36 dB SNR from the highest location ‘kaaf’ at -9.69 dB SNR to the lowest for ‘djeem’ at -16.42 dB SNR. The locations for ‘codenames’ ranged within ±1.5 dB SNR of the mean location (-13.2 dB SNR) with the highest being -14.54 dB SNR for ‘saham’ and the lowest -11.76 dB SNR for ‘fahad’. As for the ‘directions’, the mean location was -12.74 dB SNR. Locations ranged within ±2.04 dB SNR, with ‘aali’ having the highest location at -10.81 dB SNR and ‘saree’ having the lowest location at -14.89 dB SNR. As evident in Figure 3.9, all the words exhibited steep slopes except for ‘saham’ which still had a shallower slope value of 8%/Db.

**p-value for the measure of deviance and standard errors**

All words had a very good fit except for ‘khutwa’ that had a pDev value of 0.53, which is still acceptable. Standard errors for ‘djeem’ and ‘wow’ ‘saham’ and ‘saree’ were high denoting a ceiling effect.

**Confidence intervals**

Overlap existed between confidence intervals (CIs) for the ‘codenames’ and ‘directions’, but there was a significant difference between the CIs for the word ‘djeem’ [-14.3, -18.5] and the CIs for ‘kaaf’ [-6.2, -13.1] and ’ein’ [-9.5, -14.1] in the ‘letters’ list as seen clearly in the boxplots of the word location ranges in Figure 3.11.

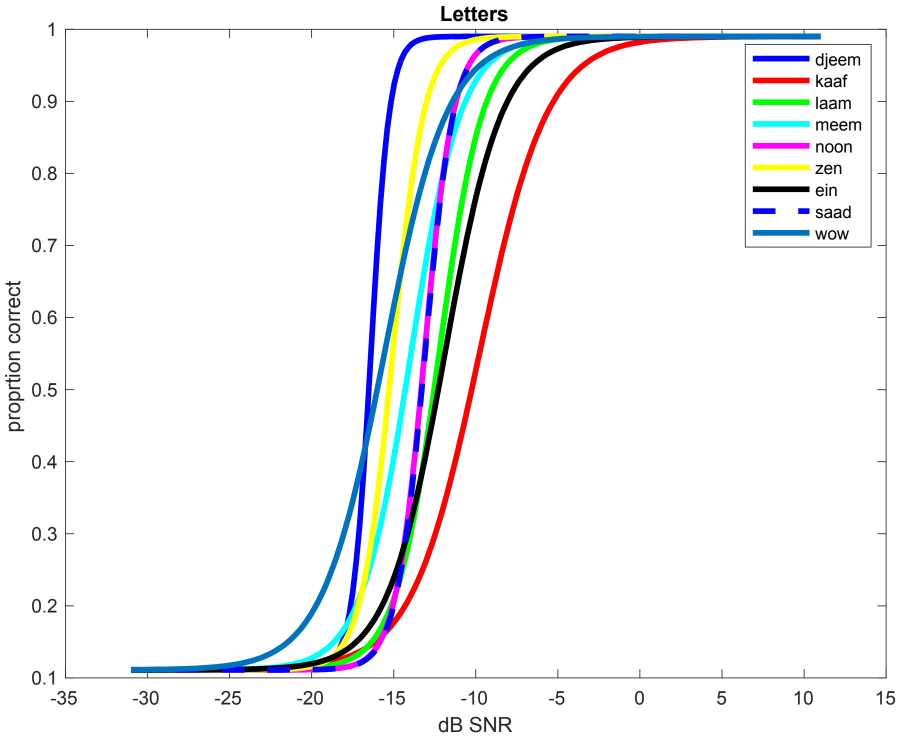


***Figure 3.8*** *Logistic functions for 'letters' in study 1.3 (obtained from the mean threshold at each SNR)*

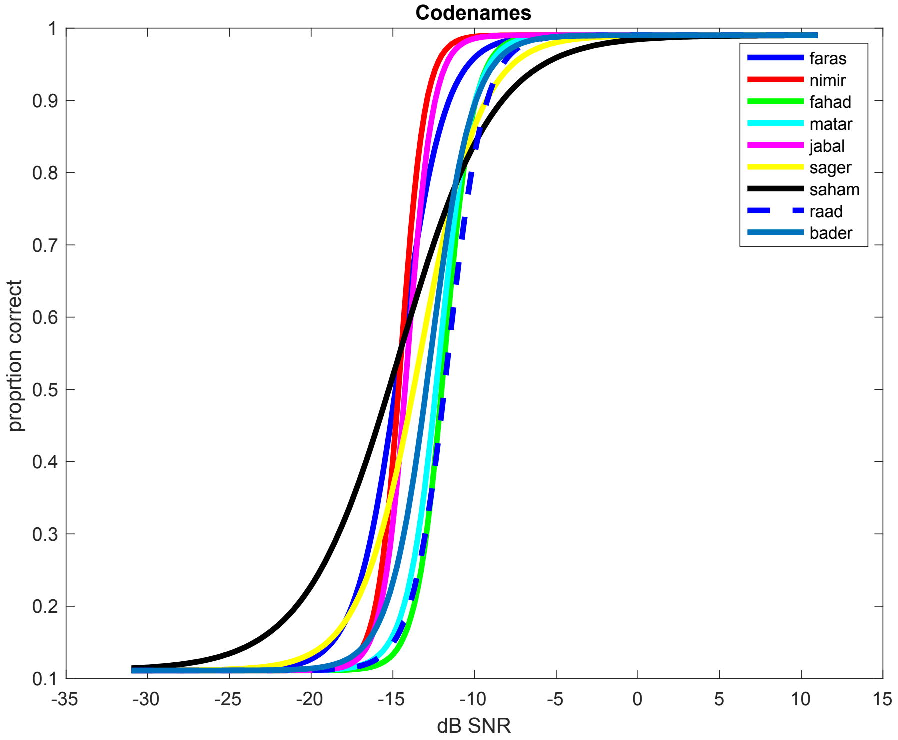


***Figure 3.9*** *Logistic functions for 'codenames' in study 1.3 (obtained from the mean threshold at each SNR)*

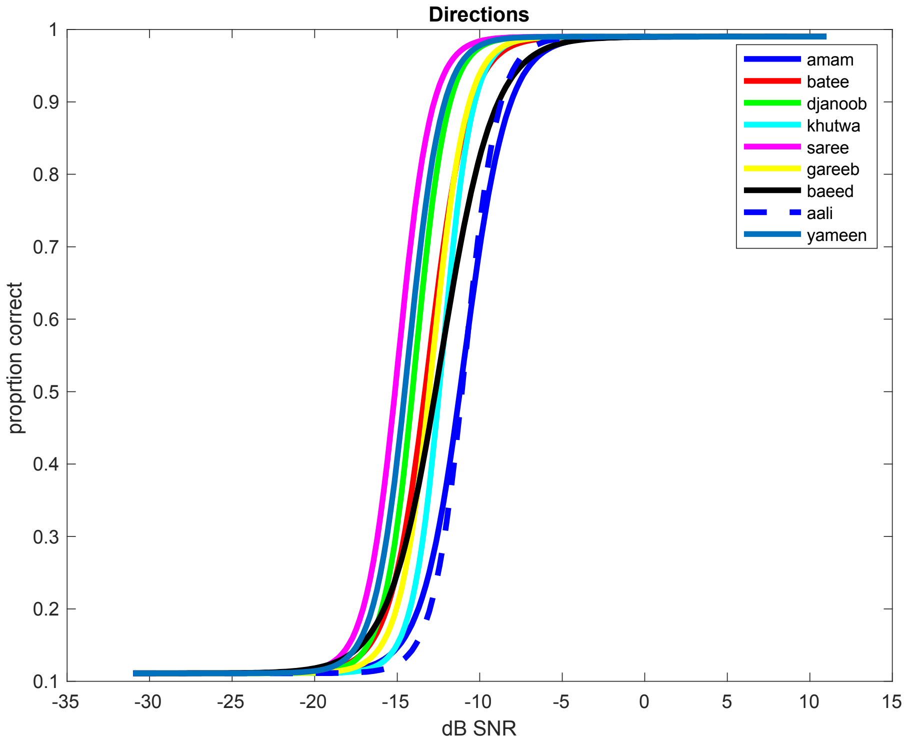


***Figure 3.10*** *Logistic functions for 'directions' in study 1.3 (obtained from the mean threshold at each SNR)*

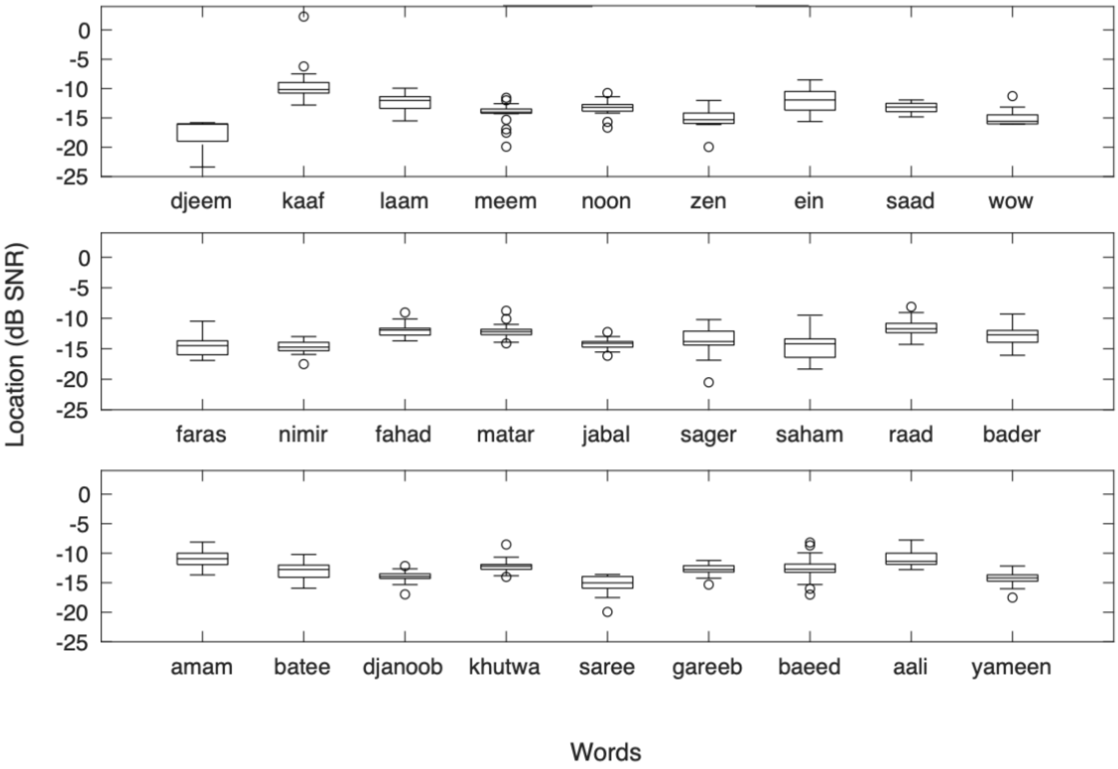


***Figure 3.11*** *Boxplots showing the range of locations in dB SNR for each list and the variation in locations for each word in study 1.3*

### 3.6.4 Discussion

Mean locations across the word lists were all within 0.5 dB of each other in session three. There was a notable change in the range between location ranges post equalization in lists two (± 2.5 to ± 1.5) and three (±4.2 to ± 2.04), but not in list one (±3.7 to ±3.4) because of ‘djeem’. Across lists the slopes were steep, apart from ‘wow’ and ‘saham’ indicating reliability of the estimated locations.

The first aim of this study was achieved. Test material was selected, developed, and tested to assess intelligibility. The second aim was partially met. Words were optimized by RMS equalizing word intelligibility but, results of retesting to assess the effect of equalization were suboptimal. The locations of the words span across a range of ± 1.5 (codenames) to ±3.4 (letters). The range of the ‘letters’ list is slightly higher than preferred, as most studies developing similar material have a range of ±0.35 to ±2.7 dB SNR in the material post-equalization ((Killion *et al.*, 2004; Smits *et al.*, 2004; Ozimek *et al.*, 2009; Kollmeier *et al.*, 2015; Leclercq, Renard and Vincent, 2018). This may be attributed to the method of testing used.

The method of constant stimuli is subject to floor and ceiling effects. The ceiling effect can be seen clearly in the words ‘djeem’,’wow’ and ‘saham’. This denotes a low upper limit to the test making it too easy. The results show that the chosen limits for the test are appropriate for all the words except these three. Taking into consideration the limitations of the method used, the absence of a general standard for acceptable variation in speech test material to be considered homogenous and the shortage of developed and validated Arabic speech material for reference, further testing to assess the intelligibility of the words is required.

The next study will assess homogeneity of the words using a modified adaptive procedure. Running the words in an adaptive procedure will eliminate the ceiling effect. It will also clarify how significant the effect of variation in the slopes and locations is on the homogeneity of the test. When consistency and homogeneity of the words have been ensured, the full test sentences will be adapted into a simple adaptive procedure. Ensuring homogeneity and repeatability will yield a tool ready to be researched as a functional occupational assessment measure of speech intelligibility in noise.

## 3.7 Study two: Optimizing intelligibility of the developed test material

The speech material developed and assessed in study one needs further investigation to ensure equal intelligibility of the words and measurement precision of the test material. As discussed, the MoCS proved to have limitations in allowing us to confidently ensure the homogeneity of the words for use in a test.

This section presents the rationale behind the method chosen for re-assessing the words post-equalization, a detailed explanation of the method and the results of the experiment using the method for assessment of the speech material post-equalization.

Study two aimed to ensure the readiness of the developed speech material to be used in a test, the main characteristic being similar intelligibility of the words in a noisy background. This was met by the following objectives:

1. Selecting and preparing the test material using an alternative method to the method of constant stimuli.
2. Selecting a suitable statistical method to express the results and comparing the results to those obtained by the previous method of constant stimuli.

The alternative method selected was a modified adaptive procedure that generated estimates of the SRT for each individual while overcoming the shortcomings of the method of constant stimuli. This method will be discussed in the following section.

### 3.7.1 Staircase adaptive procedures

In an adaptive procedure, responses in each trial are dependent on the preceding response and stimulus, unlike the MoCS. Adaptive methods are advantageous in that they are faster, eliminate floor and ceiling effects and focus on the dynamic range of the PF (Hu, Swanson and Heller, 2015).

There are many different types of adaptive procedures. For this study, a staircase procedure was chosen. Staircase procedures are considered the least complex and most flexible of the adaptive testing methodologies (Leek, 2001).

The general placement rule in the staircase procedure is based on increasing the stimulus after an incorrect response with a specified increment (step up) and decreasing it after a correct response (step down). In the simple (truncated) staircase used in this experiment, the increase up (U) and decrease (D) occur after a single response (1U/1D) as illustrated in Figure 3.12. This is the simplest form of staircase procedure which targets a 50% performance level. A consecutive series of steps in one direction up or down is labelled a *'run'*. One of the variations from the simple staircase is Wetherill and Levitt's transformed up/down method in which the difficulty is varied by varying the step size rule so that depending on the specified rule, two or three consecutive correct responses are required to increase difficulty while one incorrect response decreases difficulty. Transformed staircases with step-size rules 2D/1U, 3D/1U, and 4D/1U converge to probabilities correct of 70.7%, 79.4%, and 84.1% respectively. Other variations include Kaernbach's weighted up/down method and Garcia-Perez's transformed weighted up/down method.

Several parameters affect the adaptive staircase procedure. These parameters will be discussed in in the next section.

### 3.7.2 Test methodology

A modified adaptive staircase procedure was chosen for this experiment. The previously developed test material consists of three wordlists each containing nine word-options. One of the drawbacks of the regular adaptive procedure is that it would produce one SRT for the word set without information regarding each word individually. This does not fit the requirement of this experiment in which we are interested in the individual SRT's of the words. Performing the adaptive procedure for each word would not be very sensible as it would enable the participants to anticipate the difficulty and nature of the stimuli which would systemically bias the procedure. To overcome this, the test was modified to present all the word tracks interleaved together non-sequentially, to randomize the stimuli, overcoming the effect of focus and familiarization within the test material, thus minimizing bias, and reducing predictability (Levitt, 1971; Daia, 1995).

### 3.7.3 Parameters and analysis of the interleaved adaptive procedure

It is important to know the accuracy and precision of the developed material when the test is presented with this modified adaptive procedure. Accuracy in this context relates to the amount of systemic error or bias in the test results and is determined by the difference of the SRTs across participants from the mean population's SRT for each word. This is expressed by the SE. Precision relates to the amount of noise or randomness in the test results and will be expressed by the standard deviation (Treutwein, 1995). In this section, we consider the parameters of adaptive procedures and how they affect the efficiency, accuracy, and precision of the test, and based on these effects, the parameter values used in the experiment will be specified.

### Starting value

The starting value is the value at which the test begins. The optimum starting value is close to the threshold but in many cases the exact value of the threshold is unknown. In that case, a range of expected values must be assumed for the threshold, and a value higher than the highest value in the range is chosen, preferably a value at which a response of near 100% is guaranteed (Levitt, 1971). Based on the results of the previous experiment and our knowledge of the PFs of the words, a starting value of -6 dB SNR was chosen.

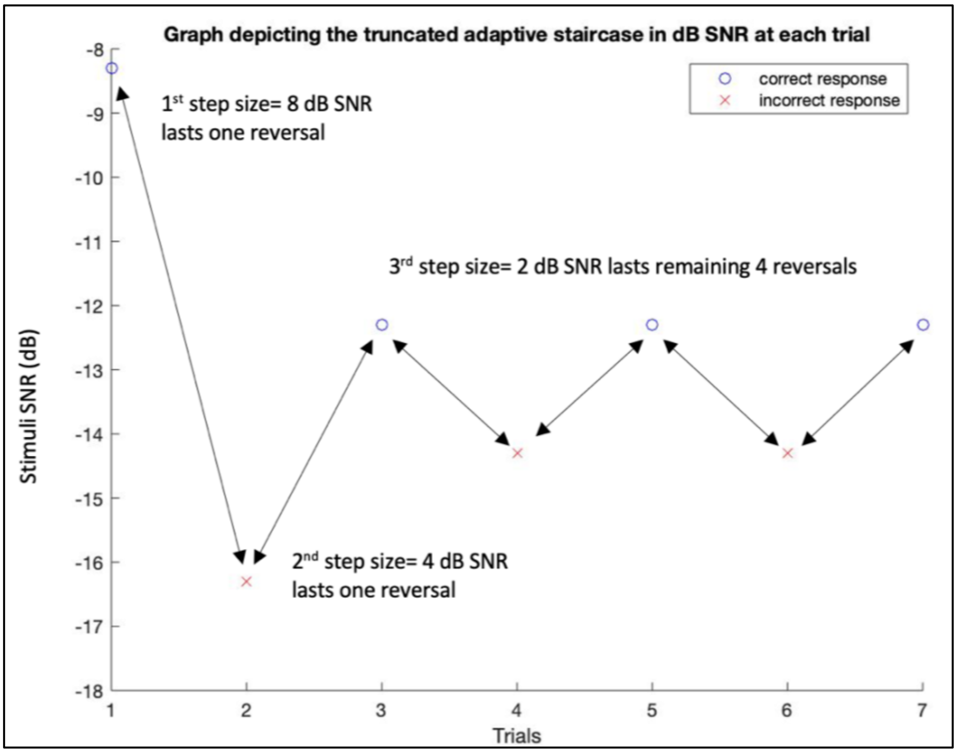


***Figure 3.12*** *Example of a truncated adaptive staircase procedure, adapted from the parameters of study 2, termınating after six reversals.*

### Step size

The size of the steps affects the speed and efficiency of convergence to the SRT. The larger the step size the quicker to converge to the SRT and vice versa. If the slopes are known or an estimate of slopes is known, then larger step sizes are more efficient for shallow slopes, and smaller step sizes are more efficient for steeper slopes. For this test, as illustrated in Figure 3.12, the chosen step sizes were 8 dB for the first reversal, 4 dB for the second reversal, and 2 dB for the remaining reversals. The larger step sizes at the beginning allow for easier presentation levels at the beginning which can quickly converge to the area of the estimated threshold.

The step size rule is the rule that guides the change in reversals. A reversal occurs when the stimulus value changes direction from the stimulus value directly preceding it. Increasing reversals increases the precision of the test and ensures sufficient data collection for threshold estimation but lengthens the test procedure. While varying the step size rule to 2D/1U OR 3D/1U targets a higher performance level, it disadvantageously lengthens the test. A truncated staircase method was used in this study to simplify the test.

**Termination rule**

The stop rule is the rule that determines the terminating point of the test procedure. It is usually based on a predetermined number of trials or reversals. Standard practice in psychophysical testing is to discard the first few reversals with the intent of eliminating the effect of increased noise from the range-finding reversals.

To balance between sufficient data collection and reasonable procedure length, six reversals were required for completion of each repeat of the test, discarding the first two reversals, resulting in a total of four scored reversals. Each wordlist was repeated twice, and repeats were averaged resulting in eight scored reversals across repeats. This was done to decrease the length of each repeat and mitigate the effects of fatigue and loss of attention. This number was chosen based on similar tests (Brand and Kollmeier, 2002; Semeraro, 2015).

### 3.7.4 Overview of the test procedure

This experiment, approved by both the University of Southampton (ERGO ref: 53345) and King Abdul-Aziz University (754-19) was conducted on 20 participants. It consisted of one session lasting approximately 90 minutes including breaks. The adaptive procedure was set-up to calculate mean scored reversals (MSR's) for each wordlist and terminate the test after eight MSR's. The MSR is a simple method developed by Wetherill (1963) used to estimate the SRT by averaging the high and low tips of all the runs. To mitigate the effects of boredom and inattention, each wordlist was repeated twice calculating the four MSR's in each repeat and averaging the repeats. The session consisted of blocks, within which a wordlist was presented, and since each wordlist was tested across two repeats, the session consisted of six blocks. Each block required approximately 10-15 minutes to complete. Participants were given the opportunity to take a break after each block, in addition to an automated break that was presented mid-block after completion of 70 trials. During the sessions, participants were sat at a table facing a computer graphical user interface (GUI). The test was administered binaurally through circumaural headphones. They listened to the recorded words preceded by a carrier phrase "ready". For the background noise, using a specific MATLAB code, SSSN was generated by digitally filtering gaussian noise to have the same LTASS of the recorded words, concatenated together.

### 3.7.5 Results

Data collection started in the beginning of March 2020. The sample size necessary to assess the differences between words was calculated based on a mean difference of 0.9 dB SNR and a SD of 1.4 dB, similar to findings of previous studies (Kollmeier *et al.*, 2015). For this study sample size calculations were performed. It was found that 20 subjects were sufficient to detect a mean difference of 0.7 dB SNR with a power of 0.8 and a significance value of 0.05. The researcher planned to recruit 25 participants to account for possible drop-offs, but due the beginning of the COVID-19 pandemic outbreak in Saudi Arabia, and the closure of universities on the 10th of March, data collection was suspended, and analysis was done on the available data collected from 20 participants. Based on the sample size calculations done for the study, the available data was sufficient to preserve the required power and effect size.

Speech recognition thresholds were obtained by calculating the MSR's. The participants' mean SRTs were averaged to obtain the mean SRT for each word and compare the words in terms of variability for each wordlist. Looking at the descriptive data for the wordlists summarized in Table 3.5, wordlist one had a mean SRT of -13.33 dB SNR (SD= 1.57, SRT range ± 2.65 dB SNR).

Wordlist two had a mean SRT of -13.19 dB SNR (SD= 2.37, SRT range ± 2.65 dB SNR) with a notably large SD for 'saham' = 6.46 dB SNR as illustrated in figure 3.13 and the largest SE across wordlists SE= 1.31.

Wordlist three had a mean SRT of -13.10 dB SNR (SD= 1.63, SRT range ± 1.83dB SNR) with a large SD for 'khutwa' = 4.17 dB SNR and SE=0.93.

There was normality for all the words as assessed by the Shapiro-Wilk test of normality except for the word's 'wow' in list one and 'khutwa' in list three. This means that the distribution of the means for these words across words is not normal. Usually, non-parametric tests are employed in cases of deviation from normality, but since the statistical test used is quite robust and the assumption was violated only for these two words due to outliers evident in Figure 3.13, parametric testing was carried out.

Values below -22 dB SNR were considered out of range and thus classified as outliers as it is highly unlikely that any word can be identified beyond this range with the chosen SNR. Based on this, eight data points from a total of 1080 data points were classified as outliers, equivalent to 0.7% of the data. Outliers were converted to the value of the highest non-outlier value in the data, a method otherwise known as Winsorization (Igo, 2010). This method of weighted modification was preferred to discarding data points especially given the small sample size.

***Table 3.5*** *Word means, SDs, SEs and 95% confidence intervals for wordlists in study 1.3.*

| **Wordlist 1** | **mean** | **SD** | **SE** | **95% confidence intervals** | |
|---|---|---|---|---|---|
| | | | | **Lower bound** | **Upper bound** |
| Djeem | -16.6 | 1.46 | 0.26 | -17.32 | -16.29 |
| Ein | -12.55 | 1.84 | 0.31 | -13.19 | -11.90 |
| Kaaf | -12.15 | 2.60 | 0.40 | -12.98 | -11.32 |
| Laam | -11.50 | 2.29 | 0.34 | -12.22 | -10.79 |
| Meem | -13.14 | 1.97 | 0.33 | -13.84 | -12.44 |
| Noon | -13.30 | 2.04 | 0.33 | -13.99 | -12.61 |
| Saad | -12.54 | 1.34 | 0.24 | -13.04 | -12.04 |
| wow | -14.36 | 2.51 | 0.48 | -15.37 | -13.36 |
| zen | -13.80 | 1.34 | 0.17 | -14.17 | -13.44 |
| **Wordlist 2** | **mean** | **SD** | **SE** | **Confidence intervals** | |
| | | | | **lower bound** | **upper bound** |
| bader | -12.39 | 2.53 | 0.46 | -13.36 | -11.43 |
| faras | -14.01 | 1.45 | 0.25 | -14.53 | -13.48 |
| fahad | -13.34 | 1.46 | 0.27 | -13.91 | -12.77 |
| jabal | -15.07 | 1.26 | 0.24 | -15.57 | -14.58 |
| matar | -12.49 | 1.65 | 0.23 | -12.98 | -12.01 |
| nimir | -14.52 | 1.93 | 0.37 | -15.30 | -13.75 |
| raad | -12.32 | 1.78 | 0.30 | -12.94 | -11.70 |
| sager | -14.77 | 2.81 | 0.53 | -15.87 | -13.70 |
| saham | -9.78 | 6.46 | 1.31 | -12.52 | -7.02 |
| **Wordlist 3** | **Mean** | **SD** | **SE** | **95% Confidence Interval** | |
| | | | | **Lower Bound** | **Upper Bound** |
| aali | -12.70 | 1.68 | 0.33 | -13.39 | -12.01 |
| amam | -11.82 | 1.62 | 0.32 | -12.48 | -11.16 |
| baeed | -12.68 | 2.85 | 0.53 | -13.79 | -11.56 |
| batee | -13.17 | 1.47 | 0.24 | -13.67 | -12.67 |
| djanoob | -15.09 | 1.36 | 0.20 | -15.51 | -14.66 |
| gareeb | -14.26 | 1.23 | 0.22 | -14.72 | -13.81 |
| khutwa | -11.44 | 4.30 | 0.93 | -13.39 | -9.48 |
| saree | -13.72 | 1.54 | 0.30 | -14.34 | -13.10 |
| yameen | -12.99 | 1.31 | 0.22 | -13.46 | -12.53 |

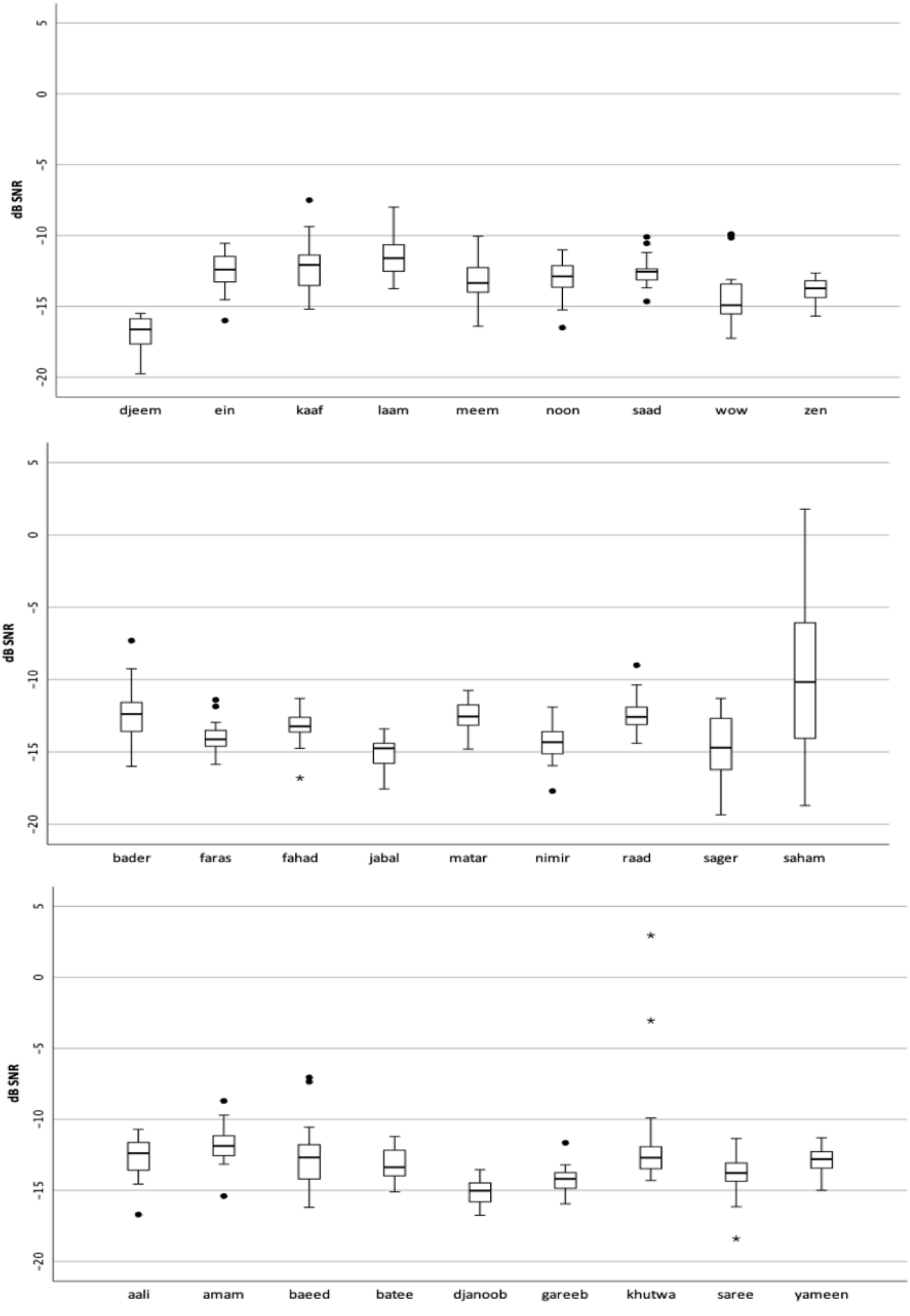


***Figure 3.13** Boxplots of the words in each wordlist for study 2. Each word is represented in the X axis and the SRT's are represented in the Y axis in dB SNR*

### 3.7.6 Statistical analysis

A two-way repeated measures ANOVA was run to determine the effect of different words over time on test performance (measured by SRT). Each wordlist was analysed separately. The results of each wordlist are reported in Table 3.6. The null hypotheses stated that:

1. All population mean SRT's are equal across words averaged across repeats.
2. All population mean SRT's are equal across repeats for all words
3. The variation in mean SRT across words does not depend on repeats

The first and second null hypotheses were rejected as a significant effect was found for the words and the repeats on test performance. The third null hypothesis was not rejected since variation in mean SRT across words was not statistically shown to depend on repeats. Having established the existence of differences between the words, it was of interest to identify the words that had differences compared to the other words.

***Table 3.6*** *Results of the two-way repeated measures ANOVA for each wordlist in study 1.3*

| wordlists | Effect of words | Effect of repeats | Effect of repeats on words |
|---|---|---|---|
| 1 | Present<br>F (8, 152) = 23.87, *p* < 0.001 | Present. F (1, 19) = 7.17<br>mean reduction of -0.66 dB SNR (95% CI, -1.19 to -0.15) *p* < 0.05 in the SRTs in repeat two | Not present.<br>F (8, 152) = 1.18, *p* = 0.06 |
| 2 | Present<br>F (8, 152) = 9.92, *p* < 0.001 | Present. F (1, 19) = 27.31<br>mean reduction of -1.10 dB SNR (95% CI, -1.55 to 0.66) *p* < 0.001 in the SRTs in repeat two | Not present.<br>F (8, 152) = 2.49, *p* = 0.13 |
| 3 | Present<br>F (8, 152) = 7.53, *p* < 0.001 | Present. F (1, 19) = 24.89<br>mean reduction of -0.65 dB SNR (95% CI, -0.93 to -0.38) *p* < 0.001 in the SRTs in repeat two | Not present.<br>F (8, 152) = 1.49, *p* = 0.17 |

Figure 3.14 illustrates the contrasts between the words in each wordlist. The pink boxes represent the pairwise comparisons for wordlist 1, the yellow boxes are the pairwise comparisons in wordlist 2, and the green boxes are the pairwise comparisons for wordlist 3.

From the figure, a significant decrease $p < 0.05$ in the mean SRT for 'djeem' is apparent compared to rest of the words. There were also significant decreases $p < 0.05$ in the mean SRT's of 'zen' and 'wow' compared to the mean SRT's of the words highlighted in the table.

In wordlist two, there were multiple significant differences between the words. In wordlist three, there was a significant decrease $p < 0.05$ in the mean SRT for 'djanoob' compared to all the words except gareeb. There were also a few significant differences between other words.

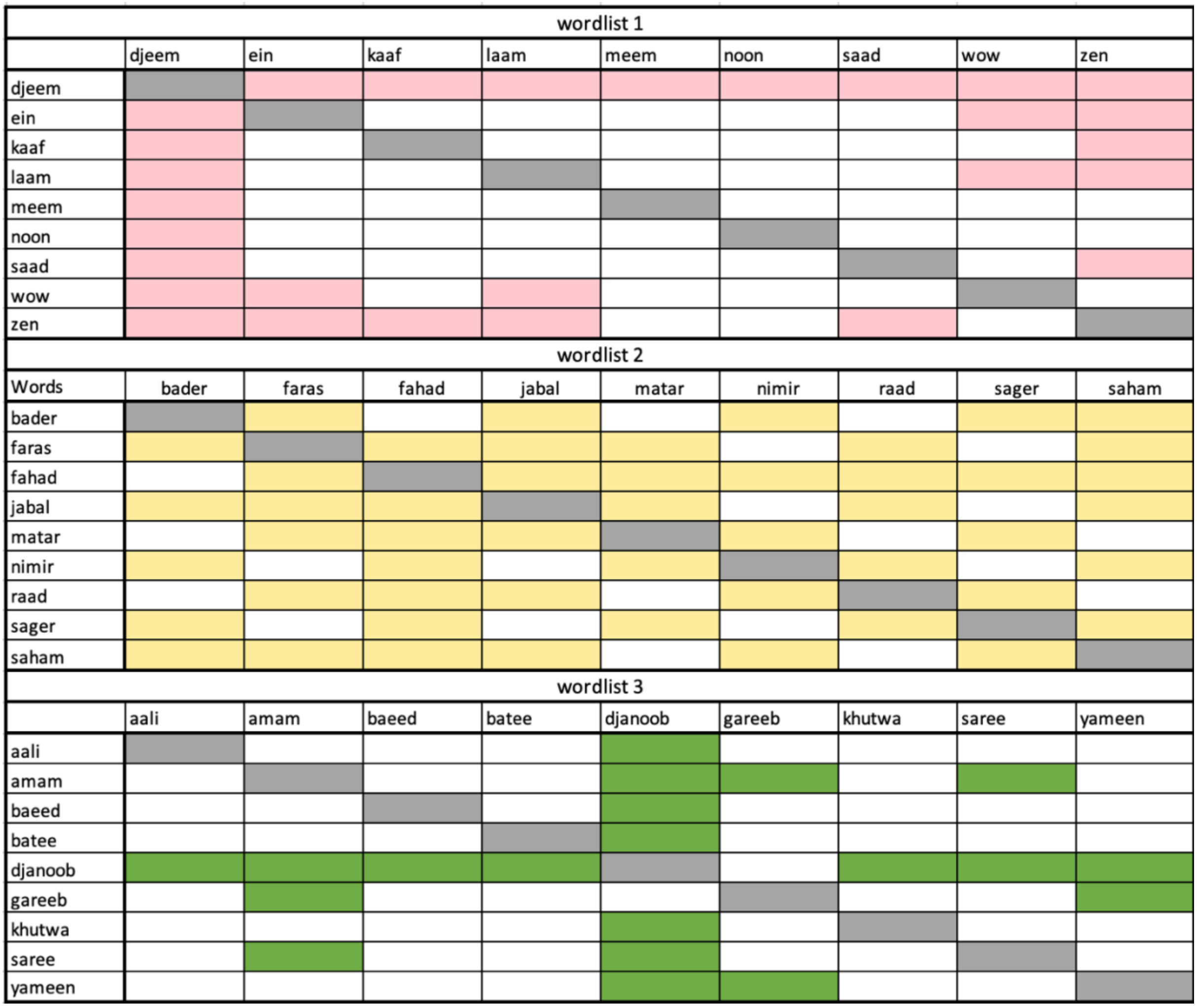
### 3.7.7 Discussion

Statistical analysis showed a significant effect of words and repeats on test performance. Regarding the effect of words, this unfortunately means that they are not equal in intelligibility, meaning that under the same noise ratio, the words differ in their recognition thresholds. Table 3.7 shows that the range of distribution of the SRT's narrowed in the adaptive procedure for lists one and three but remained the same for list two.

***Table 3.7*** *Comparison between the MoCS and the interleaving adaptive procedure in terms of mean SRT, SRT ranges and problematic words in each list*

| **wordlist** | **Method of constant stimuli** | | | **Adaptive procedure** | | |
|---|---|---|---|---|---|---|
| | Mean SRT | Range of SRT's | Problematic words | Mean SRT | Range of SRT's | Problematic words |
| 1 | - 13.42 | ± 3.4 | Wow, djeem | -13.35 | ± 2.65 | djeem |
| 2 | - 13.20 | ± 2.5 | saham | - 13.19 | ± 2.65 | saham |
| 3 | - 12.74 | ± 2.04 | | - 13.10 | ± 1.83 | khutwa |

This is possibly due to the wide range for the word 'saham' which is evident in its high SD (5.88 dB SNR). The SD for the word 'khutwa' (4.17 dB SNR) was also large compared to the remaining SDs in wordlist three. Statistical analysis showed us that the word 'djeem' is significantly easier than the rest of the words in wordlist one. Similarly, it exhibited ceiling effect in the MoCS due to its lower SRT range. Similar studies developing speech in noise tests executed the developmental phase using an adaptive procedure only. However, these studies estimated material thresholds from existing materials or from intelligibility measure computing tools (Vlaming *et al.*, 2014; Paglialonga *et al.*, 2020). For the development of this speech corpus, both studies 1 and 2 were necessary steps in the development process.



| wordlist 1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | djeem | ein | kaaf | laam | meem | noon | saad | wow | zen |
| djeem | ▨ | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 |
| ein | 🟥 | ▨ | | | | | | 🟥 | 🟥 |
| kaaf | 🟥 | | ▨ | | | | | | 🟥 |
| laam | 🟥 | | | ▨ | | | | 🟥 | 🟥 |
| meem | 🟥 | | | | ▨ | | | | |
| noon | 🟥 | | | | | ▨ | | | |
| saad | 🟥 | | | | | | ▨ | | 🟥 |
| wow | 🟥 | 🟥 | | 🟥 | | | | ▨ | |
| zen | 🟥 | 🟥 | 🟥 | 🟥 | | | 🟥 | | ▨ |

| wordlist 2 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Words | bader | faras | fahad | jabal | matar | nimir | raad | sager | saham |
| bader | ▨ | 🟨 | | 🟨 | | 🟨 | | 🟨 | 🟨 |
| faras | 🟨 | ▨ | 🟨 | 🟨 | 🟨 | | 🟨 | | 🟨 |
| fahad | | 🟨 | ▨ | 🟨 | 🟨 | 🟨 | 🟨 | 🟨 | 🟨 |
| jabal | 🟨 | 🟨 | 🟨 | ▨ | 🟨 | | 🟨 | | 🟨 |
| matar | | 🟨 | 🟨 | 🟨 | ▨ | 🟨 | | 🟨 | |
| nimir | 🟨 | | 🟨 | | 🟨 | ▨ | 🟨 | | 🟨 |
| raad | | 🟨 | 🟨 | 🟨 | | 🟨 | ▨ | 🟨 | |
| sager | 🟨 | | 🟨 | | 🟨 | | 🟨 | ▨ | 🟨 |
| saham | 🟨 | 🟨 | 🟨 | 🟨 | | 🟨 | | 🟨 | ▨ |

| wordlist 3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | aali | amam | baeed | batee | djanoob | gareeb | khutwa | saree | yameen |
| aali | ▨ | | | | 🟩 | | | | |
| amam | | ▨ | | | 🟩 | 🟩 | | 🟩 | |
| baeed | | | ▨ | | 🟩 | | | | |
| batee | | | | ▨ | 🟩 | | | | |
| djanoob | 🟩 | 🟩 | 🟩 | 🟩 | ▨ | | 🟩 | 🟩 | 🟩 |
| gareeb | | 🟩 | | | | ▨ | | | 🟩 |
| khutwa | | | | | 🟩 | | ▨ | | |
| saree | | 🟩 | | | 🟩 | | | ▨ | |
| yameen | | | | | 🟩 | 🟩 | | | ▨ |

***Figure 3.14*** *word pairwise comparison matrices for each wordlist in study 2*

The improvement in performance in the second repeat across lists ranging from 0.65 - 1.1 dB SNR is most likely due to the training effect, an effect documented to be present in most SiN tests which may observed to reach a decrease of 2 dB SNR between the first and second measurement (Brand and Kollmeier, 2002; Kollmeier *et al.*, 2015) The observed effect of training is comparable to previous studies (Theodoridis and Schoeny, 1990; Yund and Woods, 2010; Willberg *et al.*, 2020).

Based on the results, it is evident that the word ‘djeem’ from list one and ‘saham’ from list two are problematic across different testing methods either due to large variability in performance as with ‘saham’ and ‘khutwa’ or ease of the word in comparison to the other words in the word set as with ‘djeem’.  A decision was made to remove the problematic words in each list bringing the number of options in each word list down to eight. This is still an acceptable number of options in a closed set used previously in similar SiN tests (Vlaming *et al.*, 2014). In the next section, the results of piloting the modified wordlists and the effects of removing words on the word set are detailed. As for the significant difference found in the remaining words, ‘zen’ in list one, ’jabal’, ‘matar’ and ‘raad’ in list two and ‘djanoob’, ‘amam’ and ‘gareeb’ in list three, they draw attention to an important issue regarding the acceptable range of variation in homogenous speech material. The simulation study investigating this will be discussed in the next chapter.

## 3.8 Modifying and optimizing test material

The results of study two highlighted one problematic word in each list. In list one, the word ‘djeem’ was removed as statistical analysis showed it was significantly easier than the rest of the words. In lists two and three, the words ‘saham’ and ‘khutwa’ were removed respectively, as their standard deviations were too large.

The modified lists were then piloted to assess the impact of removing the problematic words on the SRT’s and to determine whether any further modification would be required. Due to time constraints and COVID-19 related difficulties in participant recruitment, a larger study was not done on the modified lists and the results for the pilot were accepted. The test was piloted on 13 participants recruited from the students and staff at King Abdul-Aziz University. The methodology is identical to that of study two with only one repeat consisting of eight scored reversals. This was done to ensure rapid and sufficient data collection.

The results of the pilot are summarized in Table 3.8. Standard deviations and ranges decreased for wordlists two and three. The range for list one remained unchanged which is not worrisome as the effect of removing the word ‘djeem’ from the first list resulted in acceptable SRT ranges and SD’s.

***Table 3.8*** *Wordlist means, SDs and SRT ranges for 13 participants on equalized 8-word lists*

| Parameter | List 1 | List 2 | List 3 |
|---|---|---|---|
| **Mean** | -13.04 | -13.74 | -13.66 |
| **Standard deviation** | 1.24 | 1.58 | 1.32 |
| **Range** | ±1.72 | ±1.2 | ±0.58 |

To assess the impact of removing problematic words, the results were contrasted with those of study two. As evident in table 3.9, the ranges improved from the MoCS and the 9-word lists. The ranges of the responses are now comparable to previous studies equalizing speech material for SiN tests.

***Table 3.9*** *Comparison of SRT ranges between MoCS equalized lists, adaptive procedure 9-word lists, and adaptive procedure 8-wordlist*

| **wordlist** | **Range of SRT's** | | |
|---|---|---|---|
| | adaptive (8 words) | adaptive (9 words) | MoCS |
| **1** | ±1.72 | ± 2.65 | ± 3.4 |
| **2** | ±1.2 | ± 2.65 | ± 2.5 |
| **3** | ±0.58 | ± 1.83 | ± 2.04 |

To assure the phonetic coverage of the wordlists, phrases and words used by the RSADF during training were compiled. Phonetic analysis was done on both the compilation and the wordlists to ensure the developed material encompassed all the phonetic sounds used in the RSADFs speech.

Analysis showed the developed wordlists cover the range of phonetic sounds used commonly and critically by the RSADF. The analyses can be found in Appendix G.

## 3.9 Conclusion

The process of developing and assessing the intelligibility of Arabic speech material was discussed in this chapter. The preference of using the method of constant stimuli for all developmental phases has shown to be questionable, due to its tediousness and floor/ceiling effects. A modified adaptive procedure has proven to be more efficient at locating the SRTs. Differences between the words in the speech corpus were observed. This is especially important to consider when using previously equalized material and making word modifications or omissions. The lists were modified, equalized, and phonetically analysed.  The resulting three 8-word lists are ready for implementation into a speech in noise sentence test. Standard deviations and range of spread of word SRT's are now comparable to previous studies with similar test materials and procedures. The speech corpus will be implemented into an adaptive procedure and named the Arabic commands in noise test (ACINT). This test will be explored as a possible measure of AFFD in chapter six.

# Chapter 4 Assessing the acceptable amount of variation in homogenous speech material using Monte Carlo simulations

## 4.1 Introduction

Study two concluded that although the adaptive procedure gave accurate results compared to the method of constant stimuli (MoCS), a degree of variation still existed between the SRTs of the words in each list. It is unclear whether this variation is within normal limits of what is considered homogenous speech test material or whether there is significant variation in the words that would adversely impact performance when implemented in a test.

Currently, no standard exists pertaining to the amount of acceptable variation in homogenous speech test material. This chapter aims to explore the question of "how close is close enough" in terms of precision and accuracy in a homogenous set of speech material. This will be done by: assessing the level of agreement between randomly chosen average SRTs across different location ranges to understand when all other test parameters are controlled for as much as possible, the amount of variation due to location ranges; assessing the extent to which varying the similarity of word locations affects precision and accuracy by running Monte Carlo simulations on several hypothetical conditions to gain knowledge regarding the acceptable amount of variation in test material that preserves its homogeneity and replicability.

## 4.2 Gap in knowledge

A crucial step in validation studies of speech material for SiN tests is ensuring the speech material is homogenous and optimizing it for equal intelligibility. This is usually done by looking at either one or more of the following measures: standard deviation of the words from the mean SRT between and across lists; range of distribution of the mean SRTs of the sentences; and resultant slopes. Most validation studies of speech material for SiN tests report standard deviations (SD) and range of distribution of obtained SRTs in the normative data as a measure of the homogeneity of the test material. Some studies include optimization of the slopes (Nielsen and Dau, 2009; Kollmeier *et al.*, 2015). The basis for the chosen values at which material is considered equalized is unclear. The Dutch digit triplet test (DTT) had SDs ranging from 1 to 1.7 dB SNR across different repeats and listening conditions (headphones or telephone) (Smits, Kapteyn and Houtgast, 2004). Brand and Kollmeier (2002) estimated SDs less than 1 dB SNR were required for the test to estimate between different listeners and conditions. The hearing in noise test (HINT) scores fell within a range of 2.4 to 2.7 dB SNR and had SDs ranging from 0.6 to 3.9 dB SNR across languages (Soli and Wong, 2008b). Nielsen and Dau (2009) developed a test that had SDs ranging between ±2 dB SNR but noted that 71% of the deviations fell within ±1 dB SNR. Equalization of speech material for the matrix test across languages was done by adjusting the distribution of the SRTs to derive steep PF slopes. Standard deviations of the SRTs ranged from 0.7 to 1.3 dB SNR across languages (Kollmeier *et al.*, 2015). The common factor between all validated studies is the lack of justification for why these ranges and values are acceptable. This study aims to address this gap in knowledge.

## 4.3 Aims and objectives

The rationale for doing Monte Carlo Simulations (MCSs) is to answer the question of 'how close is close enough?' regarding the range within which equalization is unobjectionable, and to assess the degree to which the variation within the SRTs of equalized words influences test repeatability. For this purpose, simulations using hypothetical parameters of slope and location will be run. This question will be answered by observing two scenarios; an ideal scenario fit for research purposes only and a real-world scenario modelled on the adaptive procedure utilized in the previous chapter.

This study aims to investigate and identify the acceptable amount of variation in homogenous speech test material. This will be achieved by:

1. Running a large number MCSs using several conditions and manipulating the parameters of slope and location range to create conditions extending from very low to high amounts of possible variation *(experiment 3.1)*.
2. Analysing results to assess the repeatability and variation in the test material across various location ranges and slopes and looking at the levels of agreement between randomly chosen data points across all tested conditions.
3. Running the simulations using the results of the modified adaptive procedure described in chapter 3, to evaluate the performance of the procedure and assess 'how close is close enough' in a real-world scenario *(experiment 3.2).*

## 4.4 Monte Carlo simulations

A Monte Carlo Simulation (MCS) is a computerized mathematical analytical technique that generates random data from a population, fits a pre-specified mathematical model to it, obtains the required information from the model and replicates this procedure multiple times discerning the properties and most probable outcomes of the test. It was originally designed as a risk assessment tool since this method of probability distributions is a realistic way  of evaluating uncertainty in analysed variables (Muralidhar, 2003). Monte Carlo simulations are commonly employed in psychophysical research to evaluate the accuracy, efficiency, and overall performance of test procedures (Kaernbach, 1991; Klein, 2001; Brand and Kollmeier, 2002; Doire, Brookes and Naylor, 2017; Smits, 2017; Tronstad, 2017; Garcia, Smith and Palmer, 2018), and to evaluate the goodness of fit (Maloney, 1990; Treutwein and Strasburger, 1999).

Shown in Figure 4.1 are the psychometric functions (PFs) adapted from one of the wordlists used in our previous speech material development. The figure shows the PFs of 11 words spread across a range of ±2.5 dB SNR, all exhibiting steep slopes, except for one (saham). If we were to enter the parameters of these PFs, i.e., their slopes, true SRTs, the range the SRTs are spread across, the starting value and the step sizes into the simulation code and run it, the result would be all the probable outcomes of the chosen parameters (the probability of correct and incorrect responses in each run). In this study the model from which the simulation generates the data is the PF predetermined by the slope and location range which are varied across simulations, while all other parameters are fixed.

The rationale for doing Monte Carlo Simulations is twofold. First, to assess the repeatability of the test in terms of the average SRT across multiple repeats, and the variation in the test in terms of the variation in SRT and to investigate how much varying the range across which locations are spread affects repeatability and variation. Second, to discern the acceptable range of distribution of SRTs within which equalization is considered truly equal by looking at agreement across randomly chosen results from different location ranges. For this purpose, simulations using hypothetical parameters of slope and location will be run.

The MCSs will also serve to assess the performance of an adaptive procedure using the parameters and results from the procedure described in the previous chapter.

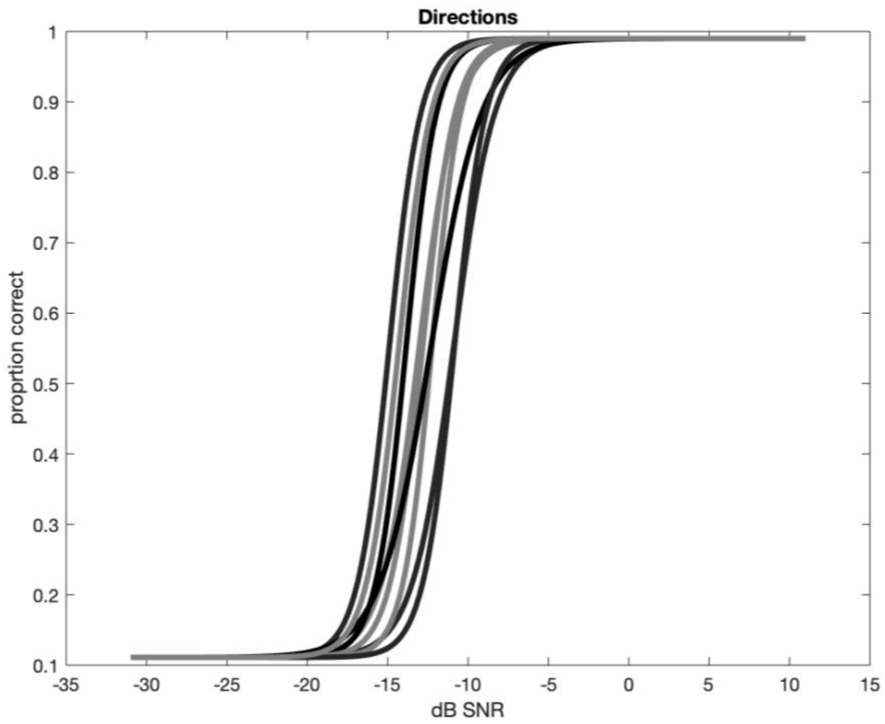


***Figure 4.1*** *Psychometric functions of wordlist 3- study 1 showing the mean proportion correct in dB SNR.*

## 4.5 Parameters contributing to variation in the test simulation

Several parameters contribute to the variation in test performance, but since the aim of the simulations is to study the effect of variation in the range of location of the words, an effort was made to optimize the remaining parameters to minimize their effect and decrease the sources of bias. Final parameter settings were chosen after review of the literature and piloting.

### 4.5.1 Slope

The slope of a PF (Figure 4.1) represents the rate of change in performance in response to change in stimulus level intensity (please refer to 3.3.3 for an explanation of parameters of a psychometric function). The steeper the slope, the more sensitive it is to smaller changes in the intensity of the presented stimulus. The size of the steps affects the speed and efficiency of convergence to the SRT. The larger the step size the quicker to converge to the SRT and vice versa. If the slopes are known or an estimate of slopes is known, then larger step sizes are more efficient for shallow slopes, and smaller step sizes are more efficient for steeper slopes as mentioned in the previous chapter.

For this study, slope values of 0.25, 0.5 and 1 were chosen to represent shallow, medium and steep slopes respectively. This corresponds to slopes of 6.25, 12.5 and 25%/dB SNR respectively.

### 4.5.2 Location

The location (true SRT) of each word has no effect on the simulation procedure in isolation, but it is important relative to the effect of starting value, the slope, and the range in which the locations of a word-set are distributed. In both studies the true SRTs were set at -12 dB SNR, and the starting levels were chosen accordingly.

### 4.5.3 Location range

This refers to the range across which the words' true locations are distributed. A location range of 3 for example, means that the true SRTs for the words in the speech material are all distributed across a range of 3 dB SNR meaning that the true SRTs of the words lie between -9- and -12-dB SNR.

### 4.5.4 Speech recognition threshold and step size rule

The percentage correct of speech recognition is determined by the type of adaptive procedure and scoring of the words. The PFs displayed in Figure 4.1 correspond to a speech recognition threshold of 50%. However, they were obtained using the method of constant stimuli. A 1d/1u procedure corresponds to an SRT 50%, meaning that the probability of a step up and step down are equal. A 2d/1u corresponds to SRT 70.7%. Inequal step size rules allow for a higher probability towards the smaller step size. In this study, a truncated staircase was chosen.

A MATLAB code written by Dr Daniel Rowan, generates the expected SRT for a specified % correct from the parameters of the PF used, and upon entering the SRT obtained from the test procedure it produces the % correct the obtained SRT represents. This code is helpful in that it allows us to gauge the effect the parameters we choose for our simulations affect the SRT, which in turn helps guide adjustments to minimize sources of variation.

### 4.5.5 Starting value

Piloting was done to understand and adjust the parameters to represent ideal conditions. One pilot compared between three starting values using three slopes and fixing the rest of the parameters. The first value was 1dB SNR above the threshold, the second was a value double the value of the threshold and the third was a value half the value of the threshold. The first two values did not show a difference in the results but the starting value that was lower than the threshold gave mean values further away from the threshold with increasingly worse performance as the slopes became shallower. According to Garcia-Perez (1998), when the starting value is above or below target, this will lead to positive (above target starting value) errors or negative (below target starting value) errors, with negative errors tending to be larger than positive errors.

The distance between the starting value and the location creates a bias and overcoming this bias requires a certain number of reversals to reach the location depending on the distance between the starting value and the step sizes. Minimizing this bias is possible by careful selection of the starting values and increasing the number of overall reversals while only scoring the last few reversals. The ideal starting point would be at the procedure's desired convergence point.

### 4.5.6 Number of variables

As this is a general study of the effect of variation on test material homogeneity, more generalizable conditions other than those specified in the researchers test conditions may be explored in future studies such as the number of variables scored in one test (one- or three-word variables per sentence) vs number of options for each variable e.g., six- or nine-word options per variable. In this study three variables were used per sentence with eight word-options per variable.

### 4.5.7 Reversals

Increasing the reversals allows for greater stabilization of the performance at the threshold. Garcia-Perez (1997) advised against using fewer than 20 reversals when conducting a truncated or 2D/1U procedure, so as not to compromise precision. While this is impractical when testing humans, it is easily applicable in simulations. One hundred reversals were initially chosen and only the last 10 were scored i.e., the mean was calculated from 10 MSR's out of 100 reversals. Further piloting clarified that 100 reversals was excessively prolonging the simulation run time. Additional piloting concluded that 50 reversals yielded results similar in precision to that gained from 100 reversals.

### 4.5.8 Type of adaptive procedure

When piloting the simulations, three methods were used and contrasted: a one wordlist adaptive procedure, a three-wordlist adaptive procedure and a three-wordlist interleaving procedure.

In the three-wordlist adaptive procedure, the guess rate was changed to 1/972 as the simulation is required to guess all three words from the different lists for a correct response. This changes the behaviour of the simulation from SRT50 to SRT80 as the required identification of three options correctly for a correct response is equivalent to a 3D/1U staircase adaptive procedure. For the interleaving methods, the guess rate was 1/9. For the reported studies, we use a three-wordlist adaptive procedure.

## 4.6 Method

### 4.6.1 Study parameters

Two studies were conducted, the first with ideal conditions and the second with more real-world conditions simulating the test developed and validated in study 2. For the remainder of the chapter, the simulations with ideal conditions will be referred to as experiment 3.1 and the real-world conditions as experiment 3.2.

**For experiment 3.1 the following parameters were chosen:**

Three slope values 0.25, 0.5 and 1. Each slope was tested at the following location ranges: 0, 3, 5, 7, 10 and 20 spread evenly. This resulted in 18 simulated conditions. Each simulated condition was run 10,000 times. The starting value was -6 dB SNR, and the guess rate was 0.0014 as three words were required to be guessed correctly for a correct response. A truncated adaptive procedure was run with the following step sizes; 4,2,1. Fifty reversals of the last step size were required to complete the test, but only the last eight reversals were scored.

**For experiment 3.2 the following parameters were chosen:**

Two slope values 0.5 and 1. Shallower slopes were not included as all the slopes in our developed test were steep. Each slope was tested at the following location ranges: 0, 1, 2, 3, 4, 5, 6 and 7 spread randomly to simulate human performance. This resulted in 16 simulated conditions. Each simulated condition was run 10,000 times. The guess rate was 0.0014. The starting value was -8 dB SNR. A truncated adaptive procedure was run with the following step sizes; 4,2,2. Twelve reversals of the last step size were required to complete the test, but only the last ten reversals were scored. Figure 4.2 illustrates one simulation run in experiment 3.2 with a location range of 7 and a slope of 1.

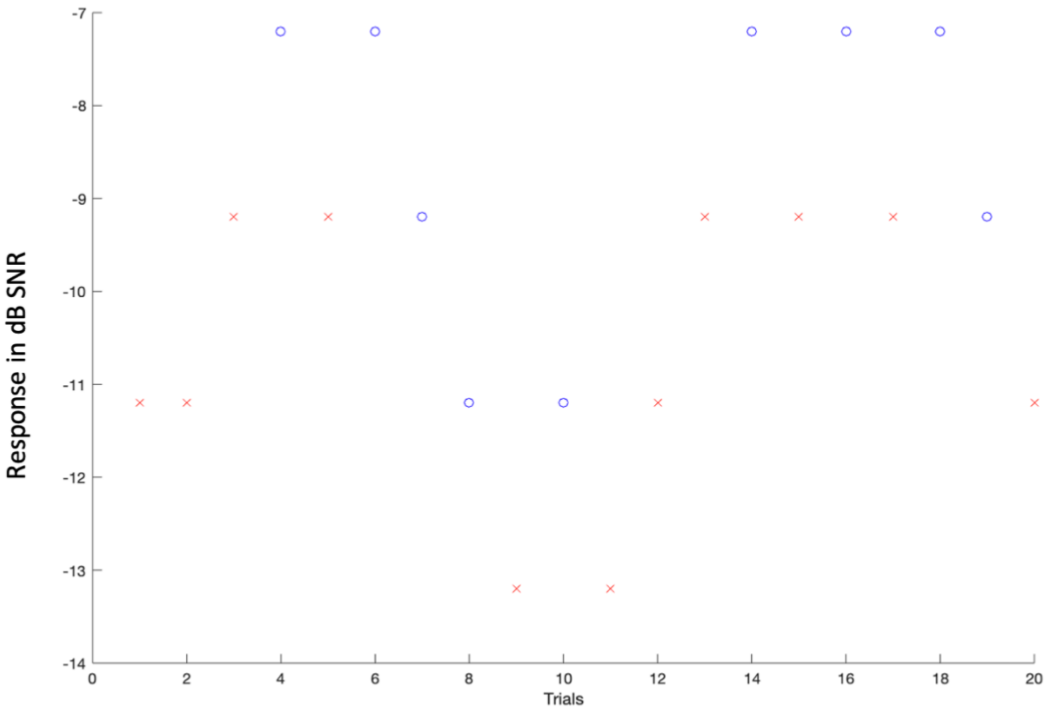


***Figure 4.2** Graph showing results from one of the simulated staircase adaptive procedures in the real-world condition used in experiment 3.2 with a location range of 7 and a steep slope. Blue circles represent correct responses and red x's represent incorrect response.*

### 4.6.2 Approach to analysis

A statistician was consulted to help choose a suitable model for data analysis. Aiming to estimate the location-to-location variability in each condition, a linear mixed model (LMM) was chosen. This statistical model considers fixed and random effects. It is preferable to traditional linear or regression models because of its handling of data cluster groups. In traditional models, the clusters are either completely separated (no pooling) or completely pooled. The LMM however, compromises between the approaches to data by partially pooling the data. This is done by multi-level modelling (Antonio and Zhang, 2014).

The model was fit to the simulation data for each experiment. The mixed model fitted contains a random effect for *location* as well as an overall mean. The model assumes the following form for $y_{ij}$, the *i*th row of the dataset, in location *j*.

$$y_{ij} = \mu + b_j + \varepsilon_{ij}$$

where *y* is the vector of observations, $\mu$ is the intercept (grand mean), $b_j \sim N(0,\sigma_b^2)$ is the effect of location *j* and $\varepsilon_{ij} \sim N(0,\sigma^2)$ is the random error.

From this model the intra-class correlation coefficient (ICC), a measure widely used in reliability analysis to assess agreement and correlation between measurements, is calculated. Intra-class correlation encompasses a set of coefficients that characterize the relationship among variables of the same class. There are many forms of ICCs used for a wide range of reliability analyses, including inter-, intra-, and test-retest reliability (Koo and Li, 2016). The ICC form is chosen based on four parameters: the model (one- or two-way), the effects (random, fixed, or mixed), the defined relationship (absolute agreement or consistency) and type (single or mean measurements) (Howell, 2010). For this study, a one-way random effect, absolute agreement, multiple measurements ICC was chosen. The ICC is the proportion of the variability in y that is due to location, this is

$$ICC = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2}.$$

Where $\sigma^2_b$ *and* $\sigma^2$ are the variance components estimated using restricted maximum likelihood (REML) method

If the location-to-location variability is a small proportion of the total variability, we can regard the test as homogeneous across locations. If this is the case, it says that measurements taken in the same location do not have anything more in common than two randomly chosen observations.

## 4.7 Results

### 4.7.1 Experiment 3.1

Standard deviations shown in Table 4.1 range from 0.49 to 1.63 dB SNR across locations and slopes. Even a very large location distribution range of 20 dB SNR had a relatively small SD of 1.17 dB SNR in the steep slopes. Up to a location range of 7dB SNR, the mean SRTs are still within 1 dB of each other.

***Table 4.1*** *Means and mean SDs (in dB SNR) for location ranges and slopes in experiment 4.1*

| **Location range** | **Shallow slope** | | **Medium slope** | | **Steep slope** | |
|---|---|---|---|---|---|---|
| | **SD** | **Mean** | **SD** | **Mean** | **SD** | **Mean** |
| 0 | 1.35 | -6.34 | 0.82 | -9.15 | 0.49 | -10.54 |
| 3 | 1.36 | -6.29 | 0.84 | -9.03 | 0.53 | -10.34 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 5 | 1.39 | -6.19 | 0.87 | -8.85 | 0.59 | -10.01 |
| 7 | 1.42 | -6.02 | 0.91 | -8.58 | 0.67 | -9.57 |
| 10 | 1.44 | -5.72 | 0.99 | -8.05 | 0.79 | -8.86 |
| 20 | 1.63 | -4.12 | 1.29 | -5.67 | 1.17 | -6.04 |

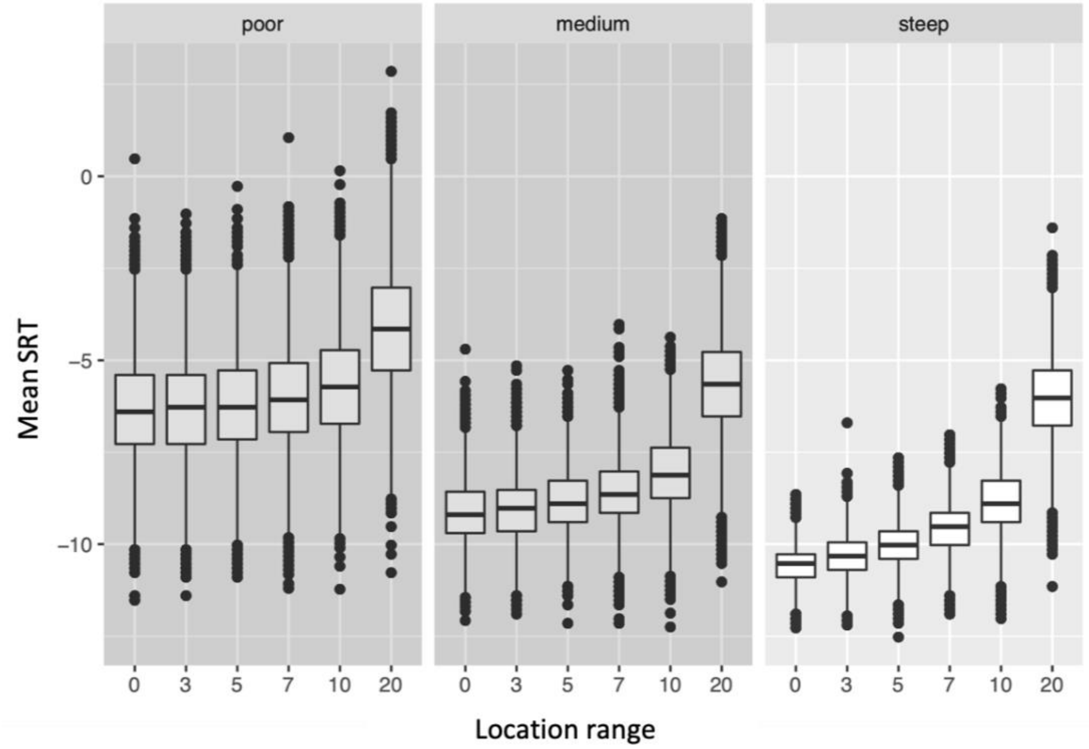


***Figure 4.3*** *Boxplots depicting the scores (mean SRTs) for 10000 runs across six location ranges in three conditions from left to right; shallow slopes (0.025), medium slopes (0.5) and steep slopes (1)*

Box plots of each location, separated out by slope are displayed in Figure 4.3. Variability due to location is apparent, especially in location range 20. The intra-class correlation coefficient reported in Table 4.2, is substantial for the medium and steep slopes, but relatively small for the shallow slope.

***Table 4.2*** *ICCs for location ranges 0,3,5,7,10 and 20 for each slope in experiment 4.1*

| **Slope** | **Intra-class correlation coefficient** |
|---|---|
| shallow | 0.26 |
| medium | 0.65 |
| steep | 0.84 |

Considering location range 20 anomalous as it resulted in a high ICC value, analysis is redone with this location range excluded (Table 4.3).

***Table 4.3*** *ICCs for location ranges 0,3,5,7 and 10 for each slope in experiment 4.1*

| **Slope** | **Intra-class correlation coefficient** |
|---|---|
| shallow | 0.03 |
| medium | 0.19 |
| steep | 0.54 |

Intraclass correlation co-efficient values for the steep slopes are still high. Excluding location ranges 10 and 20 gives the following ICCs reported in Table 4.4.

***Table 4.4*** *ICCs for location ranges 0,3,5 and 7 for each slope in experiment 4.1*

| **Slope** | **Intra-class correlation coefficient** |
|---|---|
| shallow | 0.01 |
| medium | 0.08 |
| steep | 0.35 |

The ICC of the steep slope is still greater than 0.2. Excluding location ranges 7, 10 and 20 gives the following ICCs in Table 4.5.

***Table 4.5*** *ICCs for location ranges 0,3 and 5 for each slope in experiment 4.1*

| **Slope** | **Intra-class correlation coefficient** |
|---|---|
| shallow | 0.003 |
| medium | 0.03 |
| steep | 0.20 |

### 4.7.2 Experiment 3.2

Standard deviations shown in Table 4.6 range from 0.55 to 0.94 dB SNR across locations and slopes. The mean SRTs are within 1 dB SNR for steep slopes and 1.5 dB SNR for medium slopes.

***Table 4.6*** *Means and mean SDs in (dB SNR) for location ranges and slopes in experiment 4.2*

| **Location range** | **Medium slope** | | **Steep slope** | |
|---|---|---|---|---|
| | **SD** | **Mean** | **SD** | **Mean** |
| 0 | 0.9 | -9.10 | 0.55 | -10.48 |

| Location range | Medium slope | | Steep slope | |
|---|---|---|---|---|
| | SD | Mean | SD | Mean |
| 1 | 0.86 | -9.07 | 0.58 | -10.53 |
| 2 | 0.91 | -8.80 | 0.56 | -10.16 |
| 3 | 0.93 | -8.65 | 0.57 | -10.29 |
| 4 | 0.94 | -8.43 | 0.63 | -10.19 |
| 5 | 0.95 | -8.76 | 0.63 | -10.26 |
| 6 | 0.89 | -8.48 | 0.71 | -10.42 |
| 7 | 0.91 | -7.68 | 0.63 | -9.50 |

Box plots of each location, separated out by slope are displayed in Figure 4.4. For this condition, there is less variability due to location. This is clear from the flatter box plots.

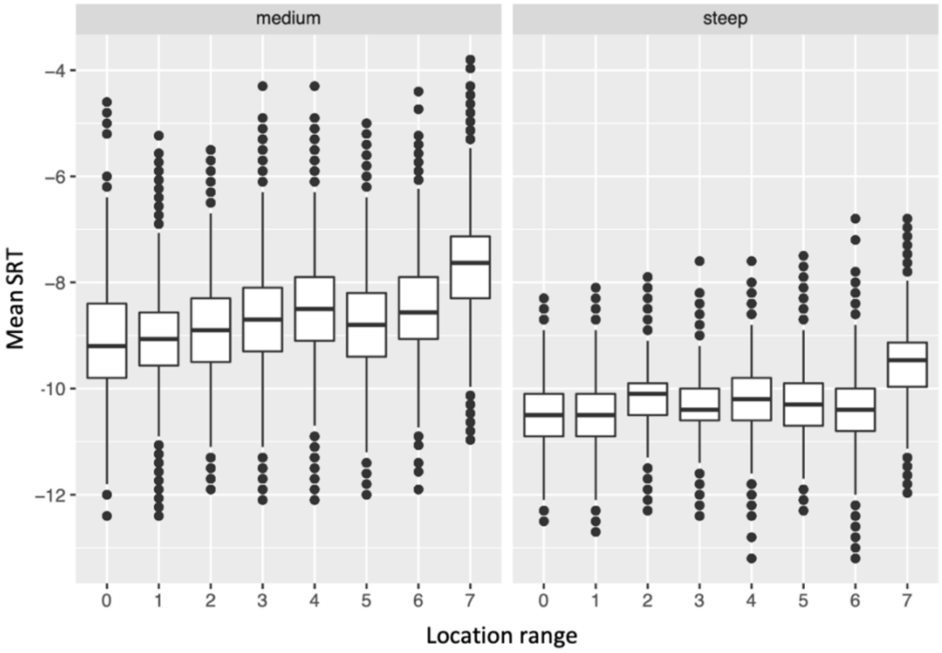


***Figure 4.4*** *Boxplots depicting the scores (mean SRTs) for 10000 runs across eight location ranges in two conditions from left to right; medium slopes (0.5) and steep slopes (1)*

The intra-class correlation coefficient for this condition shown in Table 4.7, shows that relatively little variability in score is due to location.

***Table 4.7*** *ICCs for location ranges 0,1,2,3,4,5,6 and 7 for each slope in experiment 4.2*

| Slope | Intra-class correlation coefficient |
|---|---|
| **medium** | 0.20 |
| **steep** | 0.22 |

By visual inspection of the boxplots, location range 7 is rather different in each slope condition. If location 7 is omitted, the ICC decreases substantially, corresponding to greater homogeneity amongst the remaining conditions. The ICC values for each slope, using only the remaining conditions, are given in Table 4.8.

***Table 4.8*** *ICCs for location ranges 0,1,2,3,4,5 and 6 for each slope in experiment 4.2*

| Slope | Intra-class correlation coefficient |
|---|---|
| **medium** | 0.08 |
| **steep** | 0.06 |

## 4.8 Discussion

The simulations were designed to assess the accuracy, precision, and level of agreement between locations across different ranges of distribution. The accuracy was evaluated by looking at the means across conditions, the precision by looking at the standard deviations, and the level of agreement by evaluating the ICCs. A three-variable one down/one up adaptive procedure was chosen for the simulations where the starting value and step sizes were fixed for all slopes and location ranges. This highlights the efficiency of each slope type. Shallow slopes are inefficient and less sensitive to the ability to recognize speech in noise.

For experiment 3.1, locations were distributed across a range reaching 20 dB SNR, and while the researcher is aware that this is an unacceptable range, it was done merely to highlight the deceiving effect looking at SDs and slopes only could have.

The means are used to assess the accuracy of the results. Since three variables must be identified for a correct response, the behaviour of the procedure is like a 3d/1 up adaptive procedure, creating an inequality due to the difference in step sizes (the same method was used for the matrix test across languages except that the 1d/1up adaptive procedure requires five variables in the sentence to be identified correctly for a correct response). This explains why the mean is 1.5 dB SNR away from the true SRT even at a location range of 0 with steep slopes. The means, however, are within 1 dB SNR of each other up to a location range of distribution of 7 dB SNR across all slopes in both experiments.

Intra-class correlations were calculated to assess the proportion of variation in the model explained by the variation in location and the proportion explained by random effects. In this case the smaller value ICCs indicate less variation due to location. To assure the variation due to location is low, ICC values less than 0.2 are desired to ensure homogeneity of the words across the tested location ranges.

For experiment 3.1, an acceptable ICC was obtained for shallow and medium slopes even at locations distributed across a range of 10 dB SNR. However, for the steep slope, an ICC less than 0.2 was not attained until location ranges were within 5 dB SNR. Since the test points chosen did not include a range between 5-7 dB SNR, the behaviour of the simulations in this range is unknown. Based on these results, it was decided to assess real-world conditions with locations spread up to a range of 7 dB SNR.

In more real-world conditions simulated in experiment 3.2, excellent ICCs were obtained up to a location distribution of 6 dB SNR for both medium and steep slopes.

Standard deviations were within 1 dB SNR even in location ranges subject to the effects of location variation. For instance, up to a location range of 10 dB SNR in the ideal conditions of experiment 3.1, SD ranges less than 0.8 dB SNR were seen. In experiment 3.2, all location ranges had SDs within 1 dB SNR, even though a location range of 7 dB SNR showed differences in the boxplots, and ICCs higher than desired. The SDs of the shallowest slope at a location range of 20 dB SNR was 1.63 dB SNR. These SDs are comparable to the SDs obtained in validation of speech material for the DTT (Smits, Kapteyn and Houtgast, 2004), HINT (Soli and Wong, 2008a), the multilingual matrix test (Kollmeier *et al.*, 2015) and other tests in the literature (Brand and Kollmeier, 2002; Nielsen and Dau, 2009). This suggests that depending on the slope and standard deviation may not be enough to assume homogeneity.

One study conducted similar simulations, focusing on within-subject acceptable variation in SRTs before reaching levels of significant differences, taking into consideration number of sentences, scored words, slope and location (Pedersen and Juhl, 2017) but in a context different to that of our study. Their findings showed that the factors interacted together and changes in the adaptive procedure changed the critical differences. This study focuses more on the acceptable amount of variation in speech material itself in a closed set format. The findings of Pederson and Juhl in addition to our results, caution against generalizing equalization values of one study across all types of speech studies. However, these findings may be generalised to similar types of speech test i.e., closed-set adaptive procedures. Incorporating a reliability measure such as the ICC gives a more detailed picture of the amount of agreement between locations and slopes the speech material can be spread across and still be considered homogenous. For an adaptive procedure such as the one used in validating our previously developed and equalized speech material, SRTs of the words in the word-set can be spread across a range of 6 dB SNR and still be considered equal in intelligibility.

## 4.9 Conclusion

This chapter looked at the measurements used in previously developed tests in the literature to assess homogeneity in speech material for SiN tests. Monte Carlo simulations were done to assess the acceptable range of spread of SRTs in homogenous speech material. For a closed-set adaptive procedure with three variables, the speech material can be spread across a range of 6 dB SNR and still be considered equal in intelligibility. Rather than relying on SDs and slope solely for optimization, including measures of ICC is recommended concurrent to SD and slope. The results are generalizable to similar SiN tests only and can be further expanded on by building on the work done in studying critical differences by Pedersen and Juhl (2017), and studying the effects of manipulating other parameters such as the step-size rule, the number of variables or the number of options for each variable in the test.

# Chapter 5 Assessing the feasibility of using the Arabic NEO-FFI in a military population

## 5.1 Introduction

In chapter two, factors affecting performance on an AFFD test and the gap in knowledge surrounding the effect of conscientiousness on performance in an AFFD-related hearing task were discussed. The final experiment of this PhD aims to explore this relationship. This requires using a tool to measure personality trait conscientiousness. Given the nature of military environments and training, service members may differ from civilians in personality expression. This chapter discusses the effects of military environments on employees and how this in turn may affect their personality. It will discuss the personality measure chosen for use in the study; the NEO-FFI, and the suitability of this measure for the military will be explored and contrasted to a civilian sample. This is done to assess if sufficient variation exists in the military to warrant using the NEO-FFI to estimate conscientiousness as a measure of variation in performance on a hearing task.

## 5.2 Military environments and personality expression

Military environments are rigid and highly structured. Individuals enrolled in the military are expected to behave in specific ways. One research interest was to explore the effect of personality trait conscientiousness on test performance variation in an AFFD test for the royal Saudi air defence forces (RSADF). This cannot be accomplished unless there is sufficient variation in the personalities of the test population to justify comparison of variation between individuals by personality constructs. Therefore, the feasibility of using a personality scale must first be assessed in this study.

Strong situations may place constraints on individuals' expressions of behaviour because there are predetermined and explicit expectations on how to behave. Individuals in such situations tend to behave in accordance with prescribed expectations (Mischel, 1977). The psychological pressure on individuals to engage or refrain from certain behaviours resulting from the cues given by their environment is known as situational strength (SS). Hermida (2010) defined SS as "implicit or explicit cues provided by external entities regarding the desirability of potential behaviours" (p. 122). These cues, independently or together, create environments that minimize expression of individual differences. Situational strength has four facets; clarity, consistency, constraints and consequences (Meyer, Dalal and Hermida, 2010).

Military sectors tend to be high in SS. The SS of the military environment could constrain the expression of personality.  Military supervisors were expected to be lower in consideration and higher in structure compared to their civilian counterparts working within the same rigidly defined military structure. Agreeableness, leadership and openness were constrained in an Austrian military population (McCormack and Mellor, 2002). Rolland, Parker and Stumpf (1998) examined the psychometric properties of the French translations of the NEO-PI-R and NEO-FFI on a college student sample (N=447) and a military sample of 268 male only recruits. The normative data was compared to that of the original U.S population. The military sample exhibited less variation compared to the U.S and French non-military samples. On the other hand, in a study by Campbell et al (2010) evaluating 954 US naval aviators and flight officers using the NEO-PI-R, the facet scale standard deviations exhibited a slight increase in variability in comparison to the NEO-PI-R norms. A study describing the normative data of U.S air force pilots on the NEO-PI-R found mean trait conscientiousness and standard deviations comparable to the norms of the general population, but scores of specific facets (competence, dutifulness and achievement striving) to be higher than the general population (Callister *et al.*, 1999). Huijzer *et al.* (2022) also found larger variation and higher conscientiousness trait levels in 110 Dutch special forces operators (commandos) compared to 275 male civilians. However the long version (NEO-PI-3) was used to assess the commando while the NEO-FFI was used for the control group. Similar results were found in an eight-month longitudinal study by Bech, Dammeyer and Liu (2021) comparing the Danish Navy's special warfare group  (frogmen) with civilian university students. Both groups' conscientiousness scores exhibited a slight age-related increase, with the frogmen showing a larger increase. The authors attributed this to accelerated maturity changes in conscientiousness due to the nature of training.

The construct of SS is not constant. It may vary across tasks and is subject to temporal effects, thus it is important to consider SS in this multi-level concept (Meyer, Dalal and Hermida, 2010). This may also contribute to the discrepancies in findings across studies leaving us unwilling to take findings from different settings and cultures and generalize them to our study population. This leads us to question the amount of variation in personality in a Saudi military population to justify using trait conscientiousness as a determinant of variation in AFFD task performance.

Situational strength is an important moderator in affecting expression of individual differences and personality-outcome relationships. Personality factors are more strongly related to job and task performance in situations low in SS, with environments high in the consequence facet of SS attenuating the conscientiousness-performance relationship (Meyer, Dalal and Bonaccio, 2009).

***Table 5.1*** *Studies exploring differences in variation of trait conscientiousness expression between civilian and military sectors*

| **Study** | **findings** |
|---|---|
| Rolland, 1998 | 447 French college students and 268 French military recruits were compared to the U.S normative group. Military had less variability in trait expression and highest means 38.13 ± 4.99. |
| Callister *et al.*, 1999 | Study on 1,301 U.S air force pilots. Mean and SD of conscientiousness comparable to general population norms. |
| Campbell, Ruiz and Moore, 2010 | 954 US naval aviators and flight officers. Mean conscientiousness = 46.93 ± 10.3, higher than general population norms. |
| Huijzer *et al.*, 2022 | 110 Dutch special forces commandos had higher mean conscientiousness measured by NEO-PI-3 than 275 age matched civilians measured by the NEO-FFI with a small to medium effect size *d* = 0.45 |

However, a review of the literature from 20 studies done on military populations assessing personality-outcome relationships in a military context, showed the conscientiousness-job performance relationship to be present, corrected *r*=0.35,which is not surprising given that military requirements correspond to trait conscientiousness (Darr, 2011). If testing individuals at entry level and throughout their career using the same AFFD test is required, but test performance is subject to change because of the environment, this should be considered in test development and establishment of reference values and cut-off points. the NEO-FFI will be used as a tool for measuring conscientiousness. The issue arising is ensuring the suitability of this tool in a military population. The Arabic NEO- FFI will be used to determine the variation in the study population by looking at the resulting standard deviations for conscientiousness and comparing them to a civilian control group and to existing studies. This measure will be discussed in the next section.

## 5.3 The Five-factor model

One of the most widely accepted theories of personality is the five-factor model (Costa & McCrae, 1988), in which personality is measured in five dimensions; extraversion, neuroticism, agreeableness, openness to experience and conscientiousness. Each dimension can be broken down into two main domains (DeYoung, Quilty and Peterson, 2007) and six facets. The NEO-PI-R (Costa &McCrae, 1992b) is a long-established and acknowledged personality measure consisting of 240 items. This self-report takes 35-45 minutes to complete. A shorter version, NEO-FFI, consisting of 60 items was created. It is a standard, widely used, reliable test with sufficient internal consistency (Robins et al,2001; McCrae et al, 1998; McCrae et al, 2011; Terracciano et al, 2006). In 2005, McCrae et al. obtained data from 50 cultures that had taken the test in over 20 language translations and found that the resulting coefficients indicated replicability.

Several measures exist in the literature for measuring conscientiousness in isolation. They are either extracted from a comprehensive personality measure or developed specifically to measure the trait in isolation. A comprehensive measure of personality assessing all five traits is preferable, as the isolation of one trait may affect the internal consistency or variance of the measure (Sackett *et al.*, 2017). When choosing a measure, validity, reliability, and cultural fit from the available Arabic material must be ensured.

The measures used to assess the FFT by Costa & McCrae are the NEO inventories. The NEO is a standardized widely used test validated in many languages including Arabic (Al-Ansari, 1997). The inventories consist of statements rated on a five-point Likert scale ranging from 0 (strongly disagree), 1 (disagree), 2 (neutral), 3 (agree) to 4 (strongly agree).

The NEO-PI-R is the revised long version of the NEO inventory. It consists of 240 items comprehensively assessing the five factors through 30 traits, six for each factor listed in Table 5.2, with eight statements for each trait. It takes about 45 minutes to complete.

***Table 5.2*** *Traits and facets of the five factors*

| **Trait** | **Neuroticism** | **Extraversion** | **Openness to experience** | **Agreeableness** | **Conscientiousness** |
|---|---|---|---|---|---|
| Facets | Anxiety | warmth | Fantasy | Trust | Competence |
| | Angry hostility | gregariousness | Aesthetics | Straightforwardness | Order |

| Trait | Neuroticism | Extraversion | Openness to experience | Agreeableness | Conscientiousness |
|---|---|---|---|---|---|
| | Depression | assertiveness | Feelings | Altruism | Dutifulness |
| | Self-consciousness | Excitement-seeking | actions | Compliance | Achievement-striving |
| | Impulsiveness | Activity | Ideas | Modesty | Self-discipline |
| | vulnerability | Positive emotions | values | Tenderminded-ness | deliberation |

The NEO-FFI is the short version of the NEO inventory, attached in Appendix B. It consists of 60 items assessing the five factors globally with 12 items per factor, without providing detailed information about trait facets. It is one of the most widely used short personality scales internationally (Körner *et al.*, 2015). It takes approximately 15 minutes to complete. Studies have shown it to be a reliable efficient measure of personality. A revised version: the NEO-FFI-3, was produced in 2010, with changes made to some of the sentences to improve clarity and ease of reading. This version is suitable for use by ages 12 and above (Kurtz,2020). Unfortunately, a translated and validated Arabic version of this is not available.

The effects seen in the long version are sometimes elicited in the NEO-FFI albeit with smaller effect sizes, such as the effect of sex on agreeableness, neuroticism and conscientiousness , all higher in females, the positive correlation between age and conscientiousness (Chapman, 2007), and the negative correlation between age and traits agreeableness and Openness to experience (Magalhães *et al.*, 2014; Körner *et al.*, 2015), although this is not always the case. A study assessing trait differences by sex using the NEO-FFI in an age-matched population of older adults did not find a difference in trait conscientiousness (Gogniat *et al.*, 2022), neither did Egan, Deary and Austin (2000) find a difference in conscientiousness between sexes when establishing the NEO-FFI norms in a British sample (n=1025). Note however, the female population was much smaller than the male population in their study. A study assessing the Spanish norms of the NEO-FFI also failed to elicit any effects other than an effect of sex on trait neuroticism (Manga, D., Ramos, F., Morán, 2004).

### 5.3.1 Cultural variation in personality traits

Cultures are grouped into collectivist or individualistic. In collectivist cultures, group needs may be valued over individuality and behaviour may be interdependent with their social role. Individualistic cultures place more emphasis on independence or autonomy (Timothy Church *et al.*, 2008). Saudi Arabia is considered slightly collectivist in its culture (Almutairi, Heller and Yen, 2021), even with the recent rise in individualism (Jiang, Garris and Aldamer, 2018). Openness in more traditional or collectivist cultures is expressed differently as individual pursuits are more constrained. In a Jamaican population, only traits neuroticism and conscientiousness were found to be cross-culturally generalizable (Hull *et al.*, 2010). This study highlights the reliability of certain factors especially cross-culturally. Items that have poorer cross-cultural generalizability include extraversion, agreeableness, and openness to experience, especially when using the short version; the NEO-FFI. Even though Jamaica is an English-speaking country, there could be contextual idiomatic differences in language expression.

Though there have been studies criticizing the performance and reliability in some factors due to cultural differences or factor performance and reliability issues, factor conscientiousness is the one consistently sound factor across all cultural, factor (structural validity) and reliability analyses. The factors which tend to display poor cross-cultural generalizability, varying from culture to culture, are extraversion, agreeableness, and openness to experience. In an Arab Kuwaiti population, the factors that were not cross-culturally generalizable were extraversion and openness to experience (Al-Ansari, 1997).

There are an estimated 127,000 employees in the Saudi Arabian Armed forces with the majority in the land forces and an estimated 16,000 in the Air Defence. Unfortunately, normative data on the personality of the Saudi military was not found in any accessible resources, a common problem when dealing with classified military sectors (Skoglund *et al.*, 2020).

## 5.4 Aims and objectives

This study aims to assess the NEO-FFI as a feasible measure of conscientiousness in a Saudi military population. This will be achieved through the following objectives:

1. Determining if there is sufficient variation in trait conscientiousness expression in a military population compared to a civilian population using the NEO-FFI as a measure of personality.
2. Examining the effects of sector, age, and sex on trait conscientiousness.

Based on the findings in the literature the following hypotheses are formulated:

1. Mean conscientiousness is higher in military populations.
2. Mean conscientiousness significantly increases with age
3. Mean conscientiousness is higher in females.
4. Variation of trait conscientiousness in the military population is equal to or less than civilians.

## 5.5 Method

This study was approved by the University of Southampton (ERGO II ID: 62105) and the RSADF (ID: 2825) found in Appendix F. The target population were individuals training or working in the RSADF Institute or other military sectors. Although most Saudi military recruits and employees are male, the control groups were adult male and female civilians. This is in line with the future vision of Saudi Arabia to recruit more females into military services. Recruitment has already begun, and a larger percentage of females is expected to be present in the military in the coming years.

The aim was to recruit a minimum of 100 participants from the military sector and 100 civilians. At that time, the size of the military population was not available to the researcher, and this decision was based on the widely varying numbers across studies using the NEO-FFI, ranging from 100 -1,959 (Al-Shamali,2015; Egan et al. 2000; Holden & Fekken,1994; McCrae & Costa,2004; Hrebickova et al. 2002).

The combined participant information sheet (PIS) &consent form containing a link to the online questionnaire, attached in Appendix B, was sent to the gateway keeper (commander of the RSADF training academy, the Major General) whose role was solely to distribute the form to the RSADF participants. This third-party role was required as the RSADF trainees are restricted from contact with the outside world in the first few weeks of training. Upon completion of the questionnaire, the results were automatically sent to the researcher. Distribution of the questionnaire to the control group was done by the researcher. Participants were selected by convenience sampling through WhatsApp groups and word of mouth, and links were forwarded to them.

The questionnaire included information about the study, clarification of anonymity and a tick box clarifying consent prior to answering the NEO-FFI questions. Participants rated 60 statements of the NEO-FFI using a five-point Likert scale based on how much they agreed to the statements description of themselves. Each personality construct is assessed by 12 statements. Answering all questions was mandatory for submission. In addition to the NEO-FFI questions, the following demographic data were collected: age; group 1 (18-25), group 2 (26-35), group 3 (36-45) and group 4 (>45), sex and sector; military or civilian.

## 5.6 Results

Participants were divided into four groups according to age, group 1 (18-25), group 2 (26-35), group 3 (36-45) and group 4 (>45). Throughout the following sections we will refer to them as groups 1-4. Data were gathered from 864 participants, 469 military (males = 460, females = 9) and 395 civilians (males = 143 , females = 252 ). This included 96 military participants at entry level age; group1, and 50 civilians age matched to the entry level group.

***Table 5.3*** *Mean conscientiousness scores for sample divided by sector*

| **Trait** | **Mean** | | **Standard deviation** | |
|---|---|---|---|---|
| **Conscientiousness** | **Military** | **civilian** | **Military** | **civilian** |
| | 49.22 | 48.19 | 7.19 | 6.75 |

Before focusing on the variance comparison between sectors, effects of factors sex and age must be assessed. Looking at the violin plots in Figure 5.1, a difference in the distribution of the scores by age is observed. As the participants grow older, conscientiousness scores increase.

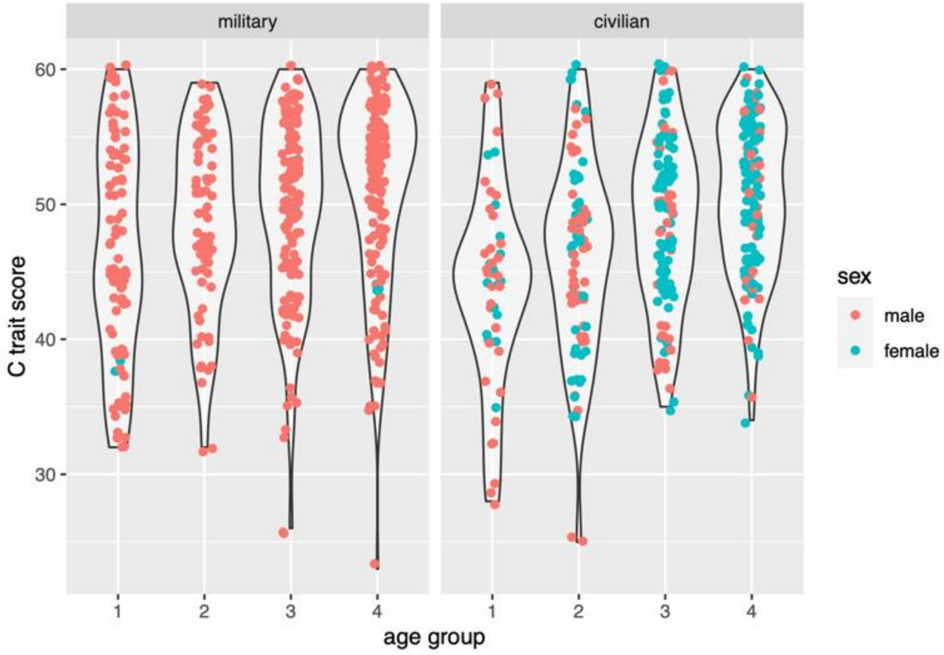


***Figure 5.1*** *Violin plots illustrating distribution of conscientiousness (C) scores by age, sector, and sex*

Looking at the histograms of score distributions across age groups in Figure 5.2, an effect of age is evident. There appears to be a ceiling effect for age with scores seen saturating at the upper limits for the older age groups. This saturation leads to reduced variability as the mean increases, causing the standard deviation to differ by 25% between groups 1 and 4.

The score has a finite maximum (60). By rescaling, it can be thought of as a proportion out of 60. Proportions generally exhibit a non-trivial mean-variance relationship: a proportion p estimated from n trials has variance p(1-p)/n. this means that relatively less sampling variability is observed when p is close to 0 or 1, and relatively more when p is close to 1/2. This violates an assumption of linear regression: that the variance of the errors should be constant. Also, any linear model fit to the data must meet the assumption of linearity. There are various ways to accommodate this violation: a variance-stabilising transformation is one straightforward approach.

The data is transformed by using a variance-stabilizing transformation (arc-sine-square-root transformation) of the response variable, approximating the data to the normal distribution. This is commonly used to transform variables whose variance depends on their value (Lin and Xu, 2020).

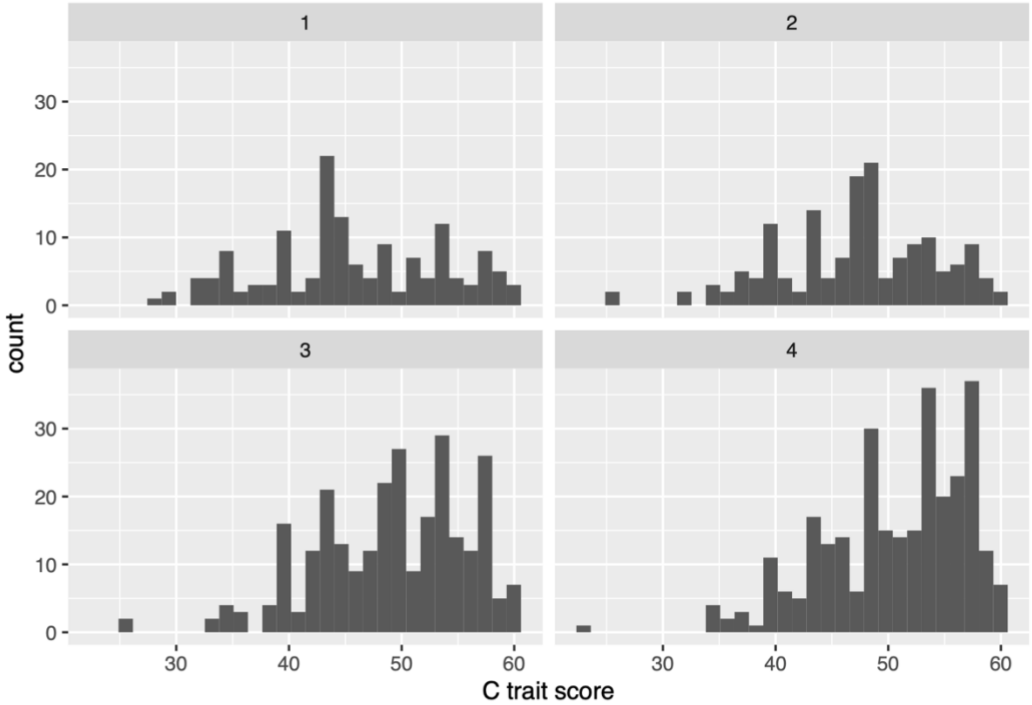


***Figure 5.2*** *Histograms showing conscientiousness (C) score distribution in each age group for both military and civilian samples combined.*

For a variable with a finite maximum (60), the appropriate transformation of the conscientiousness score, x, is

$$y = 60 \arcsin \sqrt{\frac{x}{60}}.$$

Where arcsine is the inverse sine function.

Within each age group, the standard deviations are closer to being constant as seen in Figure 5.3 (right) and Table 5.4. This approach allows use of a linear model to make estimates of the differences in means between each group, and to make statements about sampling uncertainty that are reasonable. In the presence of a mean-variance relationship, the standard errors of the mean differences would not be reliable. With a non-parametric approach, it would not be possible to control for the confounding effect of age.

***Table 5.4*** *Standard deviations of conscientiousness scores by age group before and after stabilization*

| **Age group** | **SD of conscientiousness scores** | **SD of stabilized conscientiousness scores** |
|---|---|---|
| **1** | 7.9 | 1.4 |
| **2** | 7.0 | 1.2 |
| **3** | 6.6 | 1.2 |
| **4** | 6.3 | 1.2 |

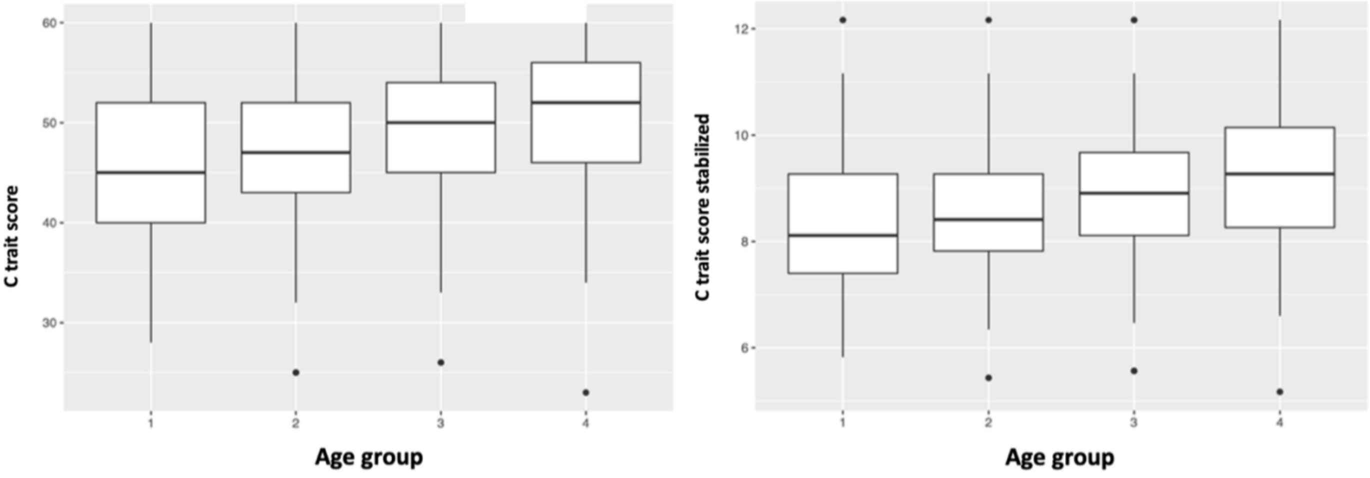


***Figure 5.3*** *Boxplots of reported (left) and stabilized (right) conscientiousness (C) scores represented by age group*

Regarding the effect of sex, average scores of males and females are similar, with a mean conscientiousness score 49.01± 6.23 for females and 48.63 ± 7.32 for males. Adjusting for age, the mean score for females is around 0.38 units lower than males; there are no substantial differences, *p*=0.43, illustrated in Figure 5.4.

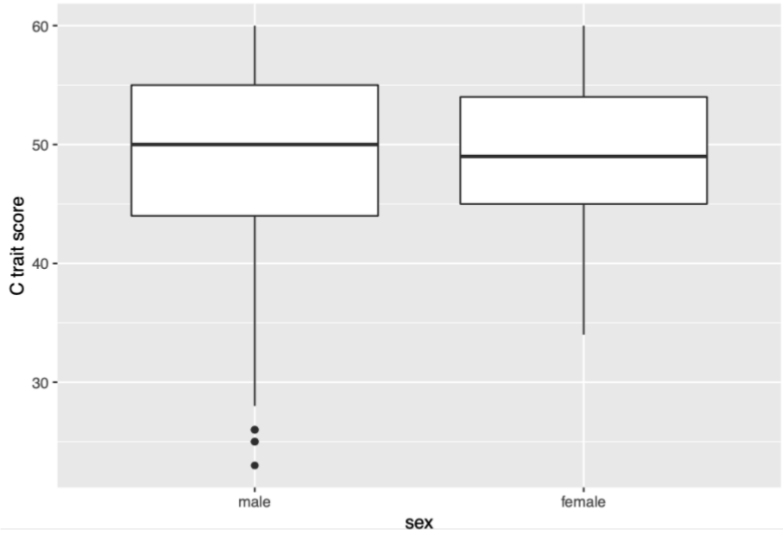


***Figure 5.4*** *Boxplots of population reported conscientiousness scores represented by sex*

Regarding the effect of sector, adjusting for age, the mean conscientiousness trait score for civilian subjects is 1.26 units lower than for military subjects; the difference is significant (p-value ~ 0.007). This is evident in Figure 5.5.

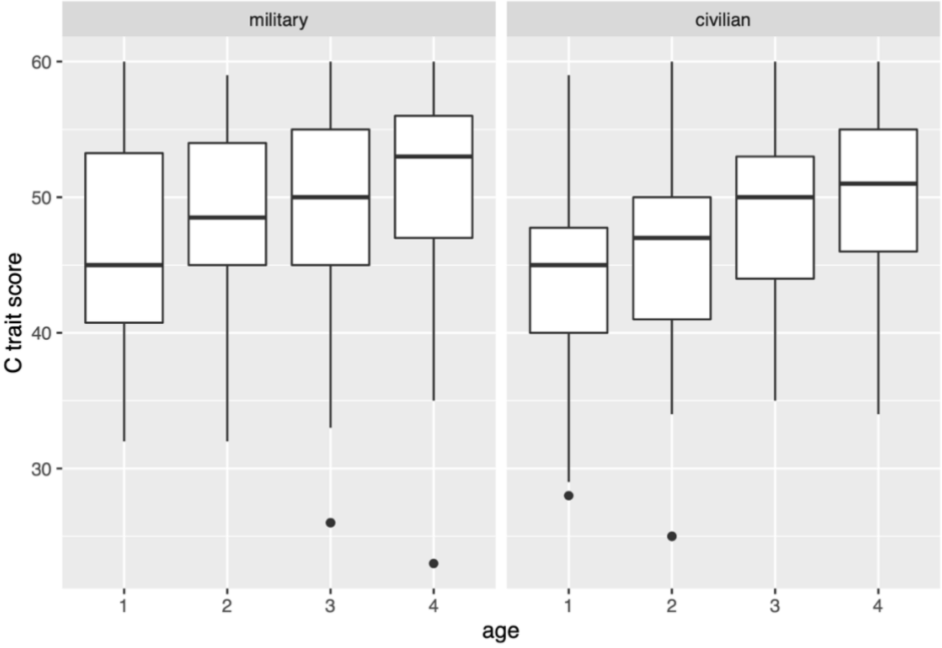


***Figure 5.5*** *Boxplots illustrating the distribution of scores across sectors, adjusted for age*

To assess the variation between sectors, a regression model is fit, and Levene's test is conducted. Levene's test is the standard test for assessing homogeneity of variances between groups. It is a robust test and appropriate to use even with violations of normality. It assesses homoscedasticity by ensuring the variance of the absolute values of the residuals; random errors, is constant (Fay, 2010). This is done by fitting a linear model to the absolute value of the residuals from the original regression model shown below. The output of the model is shown in Table 5.5

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Where *Y* is the conscientiousness trait score*, $B_0$* is the intercept*, $B_1$* is the indicator for sector *( $x_i$* is 1 if individual i is from the civilian sector and is 0 otherwise), and *$\varepsilon_i$* is the random error .

***Table 5.5*** *Output of linear model for sector effects*

| **Term** | **estimate** | **Standard error** | **statistic** | ***P* value** |
|---|---|---|---|---|
| intercept | 49.22 | 0.32 | 152.39 | 0.00 |
| Sector (civilian) | -1.03 | 0.48 | -2.16 | 0.03 |

The results of Levenes test, reported in Table 5.6, show no heterogeneity in variance between age groups, but there is evidence for some residual heterogeneity in variance between sector groups p=0.03. However, the sample size is large, so even small differences will be statistically significant.

***Table 5.6*** *Statistical output of Levenes test for the stabilized scores*

| **Term** | **Estimate** | **SD** | **Statistic** | ***p*. value** |
|---|---|---|---|---|
| Intercept | 1.11 | 0.06 | 17.34 | 0.00 |
| Sector (civilian) | -0.11 | 0.05 | -2.13 | 0.03 |
| Age group 2 | -0.13 | 0.09 | -1.53 | 0.13 |
| Age group 3 | -1.1 | 0.08 | -1.27 | 0.2 |
| Age group 4 | -0.09 | 0.08 | -1.25 | 0.21 |

By visualising the residuals from the linear model separately by sector in Figure 5.6, the differences in variability between sectors appear small, even if they are larger than expected by chance.

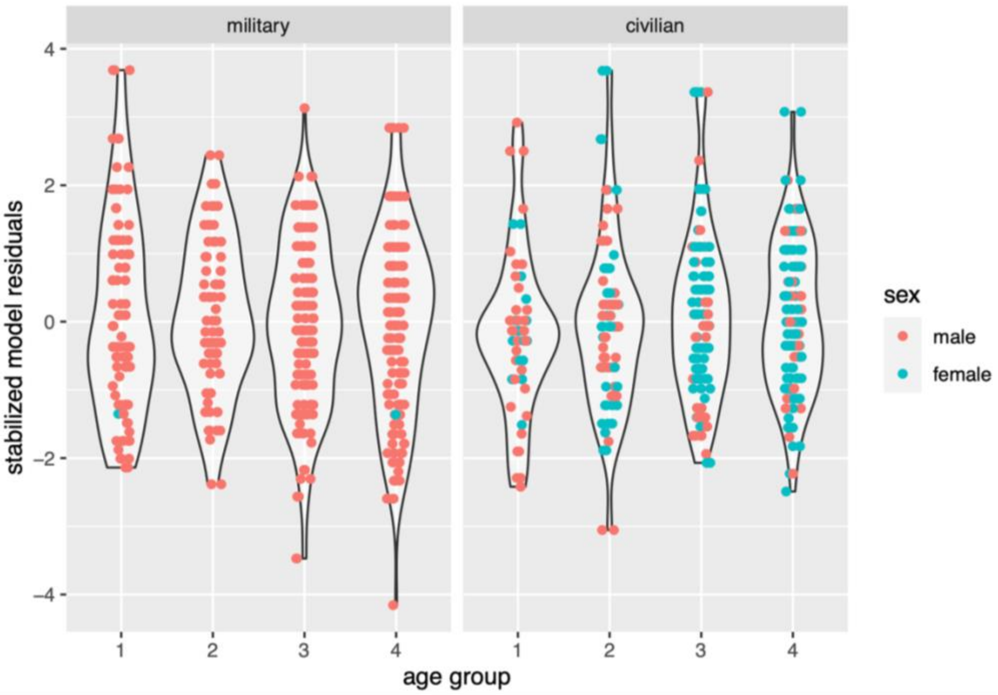


***Figure 5.6*** *Violin plots illustrating the stabilized model residuals by sector, age and sex*

To better express this, the differences in means between sectors standardized by the SD of the absolute residual in each age group shown in Table 5.7 were calculated in order to produce the effect of the difference in scores between each sector by age shown in table 5.8.

***Table 5.7*** *Mean absolute value of the residuals in each age/sector group, and their SDs*

| Age | Sector | Mean residual | Standard residual |
|---|---|---|---|
| 1 | Military | 1.18 | 0.82 |
| 1 | Civilian | 0.87 | 0.80 |
| 2 | Military | 0.92 | 0.65 |
| 2 | Civilian | 0.92 | 0.84 |
| 3 | Military | 0.98 | 0.72 |
| 3 | Civilian | 0.95 | 0.79 |
| 4 | Military | 1.04 | 0.77 |
| 4 | Civilian | 0.89 | 0.61 |

For each age group, the difference in mean absolute residual is a small compared with the standard deviation of absolute residuals in Table 5.8.

***Table 5.8*** *Effect size of difference in sector variance reported by age*

| **Age** | **Effect** |
|---|---|
| 1 | 0.37 |
| 2 | -0.01 |
| 3 | 0.04 |
| 4 | 0.22 |

## 5.7 Discussion

The main goal of this study was to determine if sufficient variation in trait conscientiousness expression existed in a Saudi military sample compared to a civilian sample to warrant using the NEO-FFI as a measure of personality. Levene's test showed a small difference in the variation of trait expression between sectors, with the military sector exhibiting larger variation in trait expression. The effects of sector, age and sex on trait conscientiousness were also examined, revealing a strong effect of age, a small effect of sector and no effect of sex on trait scores. This is in line with our first and second hypotheses postulating higher mean trait expression in military populations, and an increase of trait expression with age. Our third hypothesis was rejected as there was no difference between males and females. With regards to the amount of variation in expression of trait conscientiousness between a Saudi military and civilian sample, it was expected that trait expression would be equal to or diminished in variability in the military sample compared to civilians. This hypothesis was also rejected as variation in trait expression was found to be higher in the military sample.

An environment that constrains individuals and pressures them behave congruent to certain expectations is said to be situationally strong (Shaffer and Postlethwaite, 2013). Situational strength is also closely related to job autonomy. Job autonomy is the degree of freedom and control employees have in what, when and how work tasks and demands are executed. Decreased autonomy may weaken the conscientiousness -performance relationship (Ng, Ang and Chan, 2008). The military environment is situationally strong, and members have low job autonomy. This combination of high SS and low autonomy could affect the variation in personality expression. Military populations also tend to express personality traits differently. They are usually higher in conscientiousness and lower in neuroticism (Huijzer *et al.*, 2022).

Results are conflicting in the literature regarding the variance of expression in military populations (Rolland, Parker and Stumpf, 1998; Campbell, Ruiz and Moore, 2010). These studies are on culturally different populations. The nature of the culture; collectivist, may play a role in trait expression variance. For this study, our interest was in trait conscientiousness solely, due to its connection with task performance, and accordingly the analysis and discussion focuses only on this trait. The reason for administering a complete inventory was to preserve the measures internal consistency and stability. Using an isolated measure of the trait was not an option as a well standardized and validated measure was not available.

The resulting normative data (means and SDs) are more similar to that of the Jordanian sample (more culturally representative of a Saudi culture). Conscientiousness scores were generally higher in Arab samples (mean conscientiousness averaged  across studies (47.19 ± 6.57) (Mohaisen, 2013; Alzoubi and Alkmayseh, 2019) compared to western samples (32.88 ± 7.04) (Caruso and Cliff, 1997; Murray *et al.*, 2003; Aluja *et al.*, 2005). This might be due to collectivist nature of the population. The mean conscientiousness score for our whole sample, 48.7±6.97, was comparable to other Arabic samples.

An effect of age was found in our study, an effect documented in many studies, longitudinally and cross-sectionally, including studies by the creators of the measure, and found to be constant across cultures (McCrae *et al.*, 1999, 2004; Terracciano *et al.*, 2005; Herzberg and Brähler, 2006; Jackson *et al.*, 2009; Lüdtke, Trautwein and Husemann, 2009; Körner *et al.*, 2015) and military populations (Braun *et al.*, 1994; Bech, Dammeyer and Liu, 2021), even at the facet level (Soto *et al.*, 2011). In a longitudinal study assessing the effects of a one-year combat deployment on the stability of personality traits using the Zuckerman-Kuhlman Personality Questionnaire, a decrease in the impulsivity/sensation-seeking factor was found, mainly due to age, with the regression model predicting a 1.25 decrease in score/year for younger soldiers. It was also associated with traumatic stress exposure during deployment (Dretsch *et al.*, 2022). The study is interested in the effect of sector on the variance in trait expression. To accurately assess for sector effects, control for the effect of age was required.

After controlling for age, a minimal effect of sector was found on the variance in trait expression, 0.37, -0.01, 0.04 and 0.22 from the youngest to the eldest age group. Since variance in the military sector was higher than in civilians, this negated the concern of restricted personality expression in a Saudi military sample rendering the measure feasible for use in a military population. The increased variation in the military was also in agreement with the results of previous studies in the literature (Campbell, Ruiz and Moore, 2010; Huijzer *et al.*, 2022).

## 5.8 Conclusion

This chapter addressed the feasibility of using personality measure NEO-FFI to assess global trait conscientiousness in a Saudi military population. This was done to mitigate concerns of decreased variability in trait expression in military environments due to their situational strength.  Results showed acceptable trait expression in the military population with wider variation than the civilian population.

The differences in means and standard deviations between the two sectors were minimal and the variation of trait expression in the military sample is comparable to civilians. This allows us to explore the effect of conscientiousness level on SiN test performance in the final study. Given the relationship between conscientiousness and task performance previously discussed, the presence of a positive relationship between SiN performance on an AFFD task and level of conscientiousness is hypothesized. This will be explored and discussed in the next chapter.

# Chapter 6 Exploring performance on the developed speech in noise test under different conditions

## 6.1 Introduction

In chapter three, development of an Arabic speech corpus suitable for implementation in an AFFD SiN test was discussed. Factors affecting the performance on AFFD tests were reviewed in chapter two, and conscientiousness, a factor not widely considered yet in auditory task performance, was selected for further investigation. As a step in achieving our goal of further understanding performance in AFFD tests, this chapter will focus on exploring the newly developed Arabic commands in noise test (ACINT) as a possible measure of AFFD under different listening conditions, looking at the effect of conscientiousness on test performance and determining the test's discriminative ability.

The motivation for this research was the scarcity of Arabic language and dialect-specific SiN tests in general, particularly tests suitable for AFFD assessment. PTA remains the main, if not the only, method of AFFD testing in most sectors, including the Saudi military. However, the ability to recognize speech against a background of other sounds is an essential auditory capability. Although many studies demonstrate correlation between PTA and SiN, that does not necessarily depict or predict SiN understanding precisely. Studies emphasize the wide variation in SiN performance among individuals with normal sound detection performance by PTA (Saunders and Haggard, 1992; Saiz-Alía, Forte and Reichenbach, 2019). This could be an indication of a central problem or auditory processing issues (Musiek *et al.*, 2017). Regardless of the cause, AFFD is concerned with whether the individual can hear what is required for optimal job performance and is not concerned with hearing sensitivity alone and thus should be assessed accordingly.

The absence of functional assessment of AFFD in most FFD protocols in Saudi Arabia, including the military, is concerning and needs to be addressed. Furthermore, this study was also motivated by the need to further understand possibly overlooked factors influencing SiN test performance.

### 6.1.1 Gaps in knowledge

There is a paucity of Arabic SiN tests, and to the knowledge of the researcher none have been used for the purpose of occupational or AFFD testing. Based on this insufficiency, the ACINT was developed. The discriminative ability of the ACINT to mild sensorineural hearing loss has yet to be determined. Regarding factors that affect test performance, the relationship between conscientiousness has yet to be explored in standard and challenging auditory task settings.

## 6.2 Aims of the study

Prior to test development, a target population requiring suitable AFFD assessments was selected from a pool of suitable candidates. Initially, the RSADF training academy agreed to cooperate, and collaboration began mid 2019. Information was gathered about their roles, key communication tasks and acoustical work environments through a series of field visits to the RSADF training academy discussed in further detail in 6.3.2. Some of this work was done in parallel to test material development. Unfortunately, due to the COVID-19 pandemic and issues beyond the control of the researcher, collaboration from the RSADF was not seen through. Instead, a representative sample of secondary school students and fresh graduates was chosen for this experiment and is discussed in section 6.4.4. The goal of this study was to research the predictive ability of the developed AFFD measure on a relevant target population. This was met by achieving the following aims and objectives:

**Aim one**

The developed test, the ACINT, was adapted to be suitable for implementation in a lab simulation that best represents a specific role from the selected population’s job environment and requirements. This is discussed in section 6.3.

**Aim two**

The ACINT was tested in an experimental method to achieve the following objectives:

1. To explore the difference in performance on the ACINT under standard and task-related listening conditions in normal hearing (NH) and hearing impaired (HI) individuals.
2. To assess if performance in the task-related condition could be predicted from performance in the standard condition.
3. To explore the relationship between conscientiousness as a source of possible variation between NH individuals’ performance on the ACINT in standard and task-related conditions.
4. To examine the sensitivity of both test conditions at detecting mild SNHL.
5. To assess the repeatability of the standard test condition in NH individuals.

6. To explore the correlation between PTA and SiN performance in both conditions and compare it to existing literature.

## 6.3 Considerations for test development

Since this test was to be studied as a measure of AFFD, the framework of FFD test development, purpose, and nature of the developed test were taken into consideration.

### 6.3.1 Fitness for duty tests

There are three main types of fitness for duty (FFD) tests detailed in section 2.6. Generic predictive tests are advantageous in their broadness and ease of application. The more specific the test becomes, the harder and more expensive it is to develop, and the narrower the range of applicability is. The developed test, the ACINT, can be considered a GPT in that it has no job-related characteristics to the RSADF other than having a generally military-relevant command structure. In this chapter, we study the ACINT under two conditions, clean speech in a standard SSSN background (GPT), and speech degraded to simulate the effects of a telephone in the same standard noise background which will be considered a TPT as the modification is job characteristic. This modification was chosen for the task-related condition as it is relevant for members of the air defence. Many times, they will receive instructions through the handheld telephone receiver in their field command post. Speech through a telephone is focused on the mid-frequencies with attenuated lower and higher frequencies.

### 6.3.2 The RSADF work environment

Official approval to conduct field visits was obtained in May 2019. An introductory visit took place in July 2019 to present the research aims and requirements from the target population to the general in charge of facilitating the collaboration. A second visit was arranged in September 2019 to gain knowledge about the environment in which most tasks take place. Sound level meter (SLM) recordings of some acoustical environments were obtained, and a list of tasks was compiled. During the visit, a tour was given of the HAWK missile battery used for simulation and training including the High-power illuminating radar (HPIR) and the Battery command post (BCP) shown in Figure 6.1. Off-site meetings with SMEs were conducted to further understand the tasks and acquire more details. This information was used to adapt a suitable background for the test after development of the speech material was completed and precision was ensured.

The most important outcomes of the field visit and meetings were:

- Service members mainly operate missile batteries, namely the HAWK (Homing All the Way Killer) and the PATRIOT (Phased Array Tracking Radar to Intercept on Target) batteries.
- The BCP is the communication hub of the battery and time spent in it ranges from 15 minutes to seven hours depending on the mission.
- All internal and external communications are done through headsets or telephone receivers.
- The main tasks are listening and responding to orders in the background of steady state noise and tones.
- The average sound level in the BCP is 86 dB (A), as it is located near the HPIR which is operated via an electrical generator with an average sound level of 92 dB (A), based on measurements taken standing next to it.

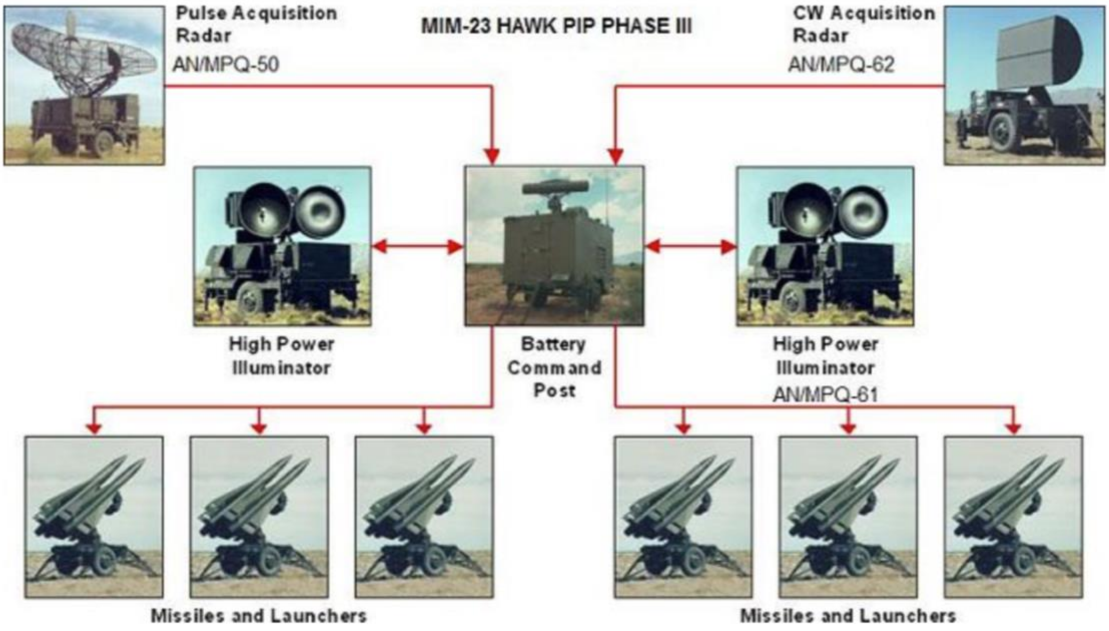
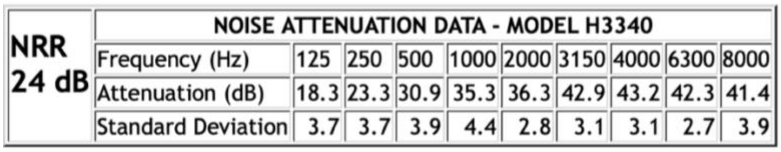


***Figure 6.1** Components of the MIM-23 Hawk PIP Phase III missile battery* (*Hawk MIM-23 low medium altitude ground to air missile technical data sheet specifications*, 2012)

### 6.3.3 Hearing protection devices (HPDs)

Hearing protective devices are recommended and used in several jobs including factory workers and military employees working in noisy environments across different sectors. They are essential for protection from noise induced hearing injuries in loud environments but often disliked by employees, because they may reduce the ability to detect and localize sounds and degrade speech intelligibility, in turn affecting situational awareness (Smalt *et al.*, 2020). They differ across models in form and function, ranging from simple occluding earplugs to dynamic filtering earmuffs (Clasing and Casali, 2014). In military employees, the criticality of the mission may be prioritized over the need to use HPDs. Training to listen while using HPDs has shown limited benefit in improving localization (Casali and Robinette, 2015). Newer models with selective frequency attenuation features may even improve listening in certain circumstances (Giguère and Berger, 2016). It is important to understand the effects specific to each model and take them into consideration when developing AFFD tests. The RSADF shared some of the HPDs used by their trainees. One of the commonly used models is the David Clark H3340 behind the head headset, shown in Figure 6.2. It has a noise reduction rating (NRR) of 24 dB. The NRR indicates the level of sound blockage provided when using the HPD. Indeed, the NRR and attenuation data provided by HPD manufacturers is that obtained from ideal lab conditions. Many service members will not have been properly educated on the correct way to wear the HPDs. For instance, simply wearing glasses will create an air gap affecting attenuation values. These are all important considerations when developing AFFD tests for HPD users.



| NRR 24 dB | NOISE ATTENUATION DATA - MODEL H3340 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Frequency (Hz) | 125 | 250 | 500 | 1000 | 2000 | 3150 | 4000 | 6300 | 8000 |
| | Attenuation (dB) | 18.3 | 23.3 | 30.9 | 35.3 | 36.3 | 42.9 | 43.2 | 42.3 | 41.4 |
| | Standard Deviation | 3.7 | 3.7 | 3.9 | 4.4 | 2.8 | 3.1 | 3.1 | 2.7 | 3.9 |

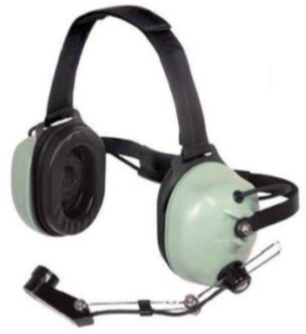


***Figure 6.2*** *Image of the David Clark H3340 Headset along with attenuation data (courtesy of the RSADF)*

### 6.3.4 The purpose of the test

The test is required to discriminate between individuals with hearing capabilities that allow them to perform their job, and individuals not meeting the requirements. The test is designed to be studied as an AFFD assessment tool for two distinct groups of the target population:

- Entry level AFFD screening, to assess individuals with no military experience prior to enrolment.

- Periodic AFFD assessments in individuals actively on duty. This encompasses levels of job experience ranging from one to approximately forty years. Different ranges of work experience may contribute considerably to improved task performance. For a detailed discussion refer to section 2.4.4.

Exploring and adapting a test for both purposes is very broad and beyond the scope of this PhD. This study focuses on the entry level test adaptation for the following reasons:

1. Entry level tests are done on individuals with no job experience. It is more feasible to use a representative population.
2. The nature of the test becomes broader as it is directed at a group with no job experience and no knowledge of the job tasks and specifics. In this case, a GPT or TPT will suffice.

The broader the test, the more applicable it is to different sectors. Because of this, the predictive ability of the test in the standard condition was to be explored and compared to a more task-related condition.

### 6.3.5 Test conditions

To compare between a generic and a more task-related test, a clean speech corpus in a background of SSSN was selected for the standard condition representing a GPT. Based on the information from the on-site visits to the RSADF, receiving commands through a telephone receiver in the BCP was found to be a frequently practised task. Due to altered cooperation from the RSADF, the frequency of using a telephone receiver for communications as opposed to using a headset is unknown. Additionally, analysis of the exact bandwidth of the telephone used was not possible. It was decided however to explore this task of listening to commands through a telephone receiver and to simulate telephone speech using Adobe Audition as a rudimentary starting point. The speech corpus was degraded to sound like telephone speech by using the telephone effect filter in Adobe Audition. This effect utilizes Fast Fourier Transform (FFT) filters, applying high and low bandpass filters that attenuate the high frequencies at around 3500 Hz and low frequencies around 300 Hz, focusing on the mid-range frequencies in the speech (Bauer, Jones and Fingscheidt, 2013). Since the BCP is situated next to the generator which emits steady machine noise, it was decided to keep the SSSN for the background noise in the task-related condition to decrease the sources of variation between test conditions.

### 6.3.6 The effect of speech degradation on test performance

An important consideration is the wider range of variation in performance in degraded or more challenging listening conditions. Higher intra-individual variation in test scores was found in tests with fluctuating noise backgrounds, spectrally degraded speech, such as noise-vocoded speech, and temporally degraded, time compressed speech. The variation in performance in studies looking at performance in degraded conditions was demonstrated to be reliable, generalizable and not attributed to measurement errors (Wagener and Brand, 2005a; Carbonell, 2017). The presence of this performance variation in NH individuals has been debated to be due to cognitive processes. The intertwinement of higher functions, attention, and factors such as personality traits is discussed in sections 2.4.7 and 2.5. The interaction of general mental ability (measured by the Watson-Glaser Critical Thinking Appraisal) with the achievement facet of conscientiousness was shown to predict task performance ($p$<0.01) in written information and customer-related problems (Perry *et al.*, 2010).

Researching performance on the developed test and conscientiousness may facilitate our understanding of the complexities involved in performing on SIN tests. This is especially important when test performance is reflective of job-task performance. Cognitive resources also play a role in listening to speech in adverse listening conditions (McLaughlin *et al.*, 2018) and fMRI studies have demonstrated that attention plays a role in enhancing degraded speech processing (Wild *et al.*, 2012).  A relationship between conscientiousness and task performance has been established in the literature (Bakker, Demerouti and Ten Brummelhuis, 2012), although it is argued whether this relationship is context specific (Moldzio *et al.*, 2021) or work-role specific (Wang, Liao and Burns, 2021). It is argued whether the conscientiousness -task performance relationship is evident in more complex tasks (Chen, Casper and Cortina, 2001; Cuttler and Graf, 2007; Minbashian, Wood and Beckmann, 2010) or more routine tasks free of ambiguity (Griffin and Hesketh, 2003; Thoresen *et al.*, 2004). The closest study to auditory task performance is a study looking at visual task performance, where 104 military gunmen and 103 civilians performed an optical illusion task (Muller-Lyer test) to assess visual misperception. The military group outperformed the civilian group and scored lower in impulsive/sensation-seeking on a personality test, a trait encompassing facets of conscientiousness (Zhang *et al.*, 2017).

Wong, Ng and Soli (2012) found performance in SSSN predictive of performance in real life noise. Similarly, we are interested in looking at the possibility of predicting the performance from changes in speech, not noise. Narrowband telephone speech requires more listening effort as the frequency limitations affect intelligibility due to missing high frequency components. This is markedly increased in noise (Pulakka *et al.*, 2012). Investigating the relationship of variation in performance in degraded listening conditions to conscientiousness is also of interest. As explained in chapter two, studies support a relationship between conscientiousness and better performance on tasks with increased levels of difficulty (Chen, Casper and Cortina, 2001; Cuttler and Graf, 2007; Lin *et al.*, 2019).

Moldzio *et al.*(2021) also argue that using a context-specific measure of personality was more predictive of the conscientiousness -performance relationship than general measures, however in this study, the sample is not from one occupational background and there is no availability of valid translated context specific-measures fitting to the context of our study.

## 6.4 Study parameters

### 6.4.1 The speech in noise test

The ACINT is a SiN test utilizing the speech material developed in study two. It is an automated closed-set response sentence in noise test. The masking noise is modified white noise, matched to the LTASS of the ACINT sentences. The sentence structure is "From (variable: letter) to (variable: codename) go (variable: directive command) now". The test sentence was compiled as follows: 500ms of silence were presented before and after each sentence to allow time for the noise ramping to the maximum presentation level prior to the speech signal beginning. This was done to allow sufficient time to avoid forward masking, which does not exceed 200ms (Moore, 2013).

The test was administered monoaurally. The test taker was required to guess the three variables in the sentence correctly for one correct response. There are eight word-options for each variable, and it was administered using a truncated adaptive procedure. The procedure terminated automatically after 12 reversals. As discussed in section 3.7.3, this termination rule was chosen after consideration of sufficient reversals for stabilization of test performance based on the results of study 2 and previous similar studies (Semeraro, 2015).

### 6.4.2 Repeatability and learning effects

Speech-in-noise tests are subject to perceptual learning effects previously discussed in 2.4.3. It is important to understand the amount and magnitude of learning to overcome it, so it does not confound the assessment of repeatability. Recommendations are available regarding the appropriate amount of training to overcome this effect, yet long training protocols could contribute to boredom, affecting the test-takers attention in the actual test. This is especially relevant in the case of our experiment where there were two test conditions. A decision was made to administer one 18-sentence training list for each test condition, as training from one condition is not transferable to another (Peelle and Wingfield, 2005; Schlueter et al., 2016).

### 6.4.3 Test population

To assess the test's ability to differentiate between NH and HI individuals, both NH and HI participants were recruited.

**Normal hearing group**

Since a main requirement of the RSADF was an entry level AFFD hearing test, it was preferable the test population have no effects of military training on their expression of conscientiousness to be as representative as possible of an entry level candidate. Therefore, a representative civilian sample in the age range that would enrol in the military was chosen. The target population consisted mainly of secondary school students and fresh graduates with ages ranging from 16-23 years.

One of the aims was to explore the effect of conscientiousness on performance. For convenience, this was to be done on the NH sample only.  As discussed in chapter five, there are differences between military and civilian sectors in terms of mean trait expression even after one year of training. Participants were divided into two groups based on their level of conscientiousness as identified by the results of the NEO-FFI questionnaire. It is common practice to choose the cut-off points based on the standard deviation. This was not done since the difference we are trying to detect is more subtle. Also, as seen in the histograms of mean conscientiousness score distribution in Figure 6.3, the distribution is not normal and there is saturation at the upper limit of the scores. For these reasons, the cut-off points were placed below and above the confidence intervals for mean conscientiousness obtained from our feasibility study. Participants below the lower cut-off point would be placed in the low conscientiousness group and those above the higher cut-off point would be placed in the high conscientiousness group.  Participants scoring in the range between the cut-off points were not recruited.

Based on the means and confidence intervals in Table 6.1, cut-off points chosen were values less than 42 for the low conscientiousness group and higher than 46 for the high conscientiousness group.

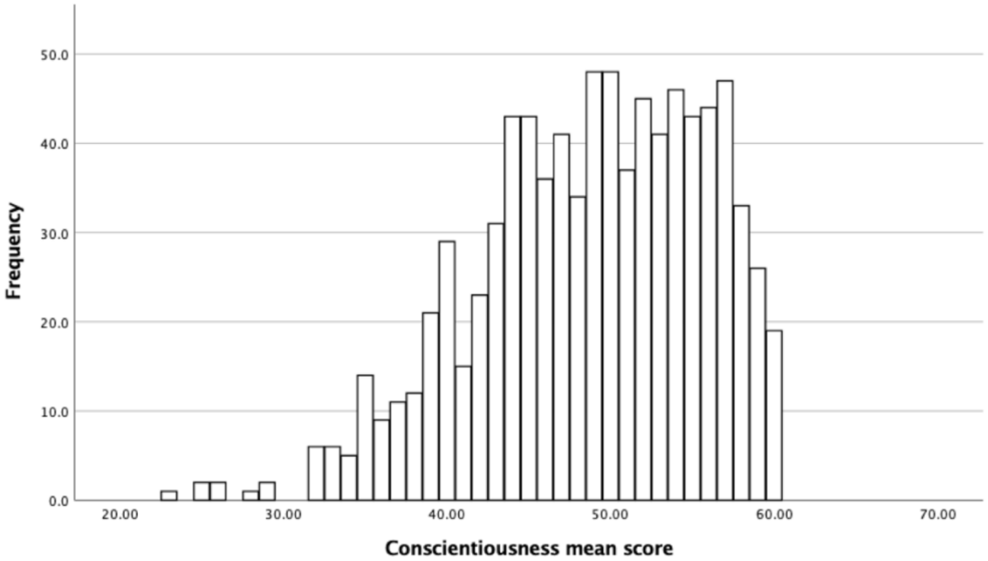


***Figure 6.3*** *Histogram of conscientiousness mean scores in study four, sample (n=864).*

***Table 6.1*** *Conscientiousness means and confidence intervals from the NEO-FFI personality questionnaire used in study four.*

| **Sample** | **Mean** | **Confidence intervals** | |
|---|---|---|---|
| | | **Lower bound** | **Upper bound** |
| Whole (n=864) | 48.74 | 48.28 | 49.21 |
| Civilians 18-25 (n=50) | 44.10 | 41.99 | 46.17 |
| Both sectors 18-25 (n=146) | 45.77 | 44.48 | 47.10 |

**Hearing impaired group**

To ensure the tests sensitivity to mild SNHL, it must detect subtle differences in hearing. Individuals with single-sided or bilateral mild sensorineural hearing loss at any frequency as identified by air (20 > *hearing threshold* ≤ 40 dB HL at 0.25-8 kHz) and bone (hearing threshold < 20 dB HL at 0.5-4 kHz) conduction PTA were recruited. Due to the difficulty of finding age-matched individuals with this type of hearing loss, the age range was chosen to be between 18-55 years of age. Regarding age related cognitive changes, performance on simple attentional tasks is maintained up to the age of 80 in normal individuals and immediate memory recall is stable even in older age in tasks not exceeding primary storage capacity (approximately six items), as is the case with the test administered in this study (three items) (Murman, 2015). Due to the limited number of individuals in this age range presenting to the clinic with the specified type and degree of HL, the HI participants were not included in the conscientiousness testing, so as not to exclude the already limited number of participants according to conscientiousness cut-off points.

Supra-threshold speech recognition, especially in noise, differs from peripheral hearing sensitivity in that it relies on higher functions distinct to those adapted in detecting low-level stimuli such as pure-tones in quiet (Carney, 2018). However, for lack of a validated widely used, Arabic test similar test to the ACINT, PTA was used to define NH- and HI populations. For a discussion on the relationship between PTA and SiN recognition please refer to section 2.3.2.

### 6.4.4 Sample size selection

The test is new, and a valid Arabic SiN test with a similar design or structure is unavailable. Three tests were considered to define the standard deviation for our population and determine the sample size: the Pediatric Arabic Auditory Speech Test (PAAST); the British English version of the CRM; and the multilingual Matrix test. The PAAST, a children's SiN test developed based on the McCormick test, was considered as it is an Arabic SiN test with a closed format. The nature of the CRM is close to that of our test. The final consideration, the Matrix test, was chosen because it is structured similarly to our developed test albeit with more variables to identify and is validated in many languages. To err on the side of caution, the highest SD; for the Polish version of the matrix test; 1.3 dB SNR was chosen. A mean difference of 2 dB between groups was specified to eliminate noise. Based on the power calculations, 20 participants were required in each group to achieve a power of 0.90.

## 6.5 Method

### 6.5.1 Subjects

Normal hearing (n= 54 (16 male), low conscientiousness = 17, high conscientiousness = 37) and HI individuals (n=20, 11 female) participated in the study. Most NH participants were tested in quiet classrooms at their schools. The remainder of the NH participants were tested in a quiet room in the audiology unit at King Abdul-Aziz university hospital, as were the HI participants. None of the HI participants used hearing aids. Ethical approval was obtained from the University of Southampton (ERGO ref: 67665.A2), King Abdul-Aziz university (REC reference no. 55-22) and the ministry of education (D3/1/38) found in Appendix E.

For the NH sample, through arrangement with the ministry of education, selected schools were approached, and questionnaires were distributed to the students' emails through the school coordinator. Consent forms were obtained from the students if they were above 18 or from their guardian if they are under 18. The questionnaire consists of 60 simple statements that require approximately 10-15 minutes to complete. Upon submission of the answers, they were directly sent to the researcher. Participants who met the criteria based on the questionnaire results were informed of the speech test appointment by the school coordinator. They were made aware that participation was optional.

### 6.5.2 Test procedure

Testing was conducted in one session. Participants attending the session had their hearing screened by a portable audiometer (Interacoustics AS608e). During screening, pure tones were presented monoaurally to the individual at the following frequencies: 0.5, 1, 2, 4, 6 and 8 kHz using a portable audiometer. The maximum presentation level of the pure tones was limited to 30 dB HL. If the individual did not respond at 30 dB HL, screening was terminated, and the participant was excluded from the study. The better performing ear was selected. They then listened monoaurally through headphones to the sentences without background noise and did 18 trials of each condition to familiarize themselves with the test.

The SiN test was then administered four times; two listening conditions, with two repeats for each condition. The sequence of the conditions was randomized for each participant using a randomizing application to overcome order effects. Each test required approximately five minutes to complete. The average session duration was 45 minutes including ensuring consent, explaining the procedure, screening, training, and breaks.

Participants listened to the ACINT sentences monoaurally over headphones, in the presence of a binaural SSSN noise masker. The stimuli were manipulated using MATLAB code (version 2019b) on a computer to control the various parameters being tested. Participants responded to the sentences using the GUI in Figure 6.4 and selecting from the available options with a mouse. Responses were automatically recorded on the computer. If the participant was unsure, they were instructed to guess. Since Co-ViD-19 restrictions were still required at the time, Saudi Ministry of Health guidance on social distancing, face-covering, and sanitizing equipment was followed (Ministry of Health, 2021).

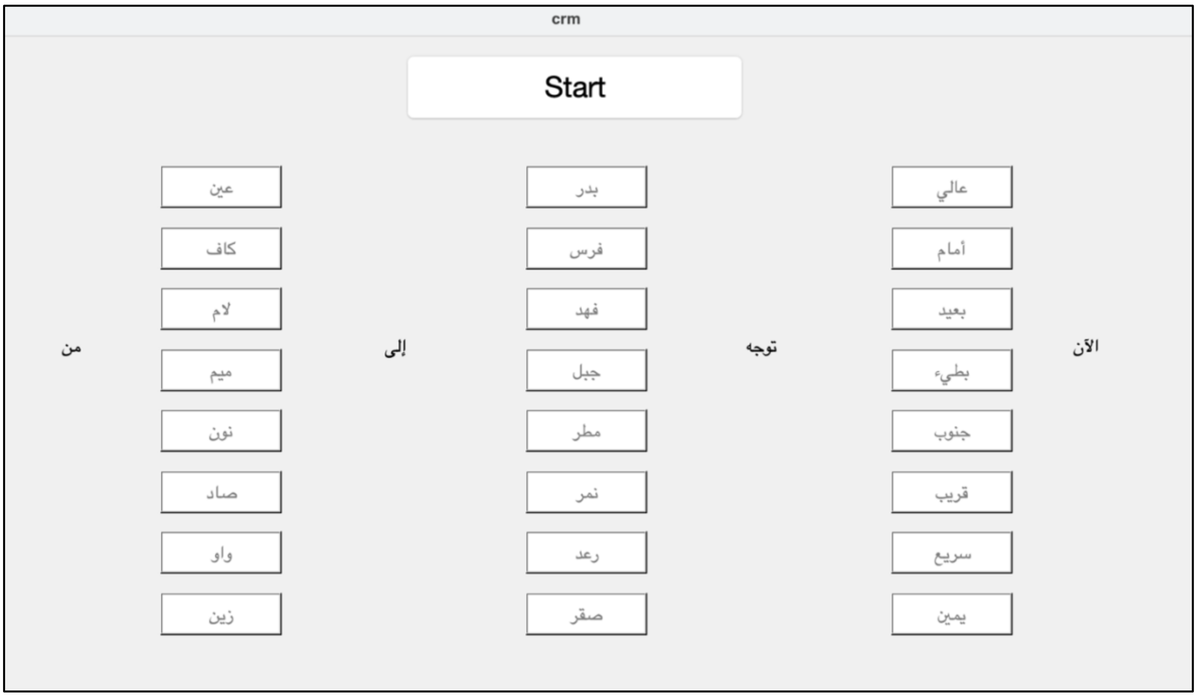


***Figure 6.4*** *Graphical user interface of the ACINT*

The noise exposure levels did not exceed the sound level which defines an unusual experiment as outlined in The ISVR Report 808- info for noise and vibration ethics. The noise exposure calculation is based on both the target ACINT sentence and the masker never exceeding 70 dB (A) independently, and so when both the target and masker are presented together the highest maximum combined level was approximately 73 dB (A). The exposure duration is based on the participant having a maximum listening time of 2 hours in any 24 hours period.

Calibration of the stimulus was measured through the headphones used for testing (Sennheiser 650 circumaural headphones) which were PAT tested. The stimulus used for calibration was speech shaped noise which has the same frequency shaping as the ACINT sentence stimuli. The headphone output was measured using an ear simulator and through a calibrated sound level meter. The objective calibration was done at the beginning of the experimental period and then weekly thereafter. Subjective listening checks were carried out at the start of each test session where the researcher listened to all the test stimuli over the headphones to check that the levels and quality sound correct.

For the HI participants, individuals attending the audiology clinic diagnosed with mild sensorineural hearing loss as identified by PTA at KAAUH were approached and asked to participate. The test session was conducted in a quiet room in the audiology department in the same fashion as that for the NH participants.

## 6.6 Results

Data were gathered from 54 NH individuals (mean age = 17.8) and 20 individuals (mean age = 37.4 years) with mild SNHL in one or both ears as identified by PTA. For the HI participants, in the chosen ears, the mean PTA configuration at frequencies 0.5, 1, 2,4 and 8 were 13.3, 14.8, 17.3, 22 and 30 dB HL respectively. Figure 6.5 shows participant demographics. One of the HI participants did not complete the second measurement of the task-related condition.

The scatterplots in Figure 6.6 illustrate the variability in the two measurements in both the NH and HI groups in both test conditions. The substantial variability in the performance of the HI group is evident, especially in the task-related condition. The change in scores between the first and second measurement is illustrated more clearly in the boxplots in Figure 6.7.

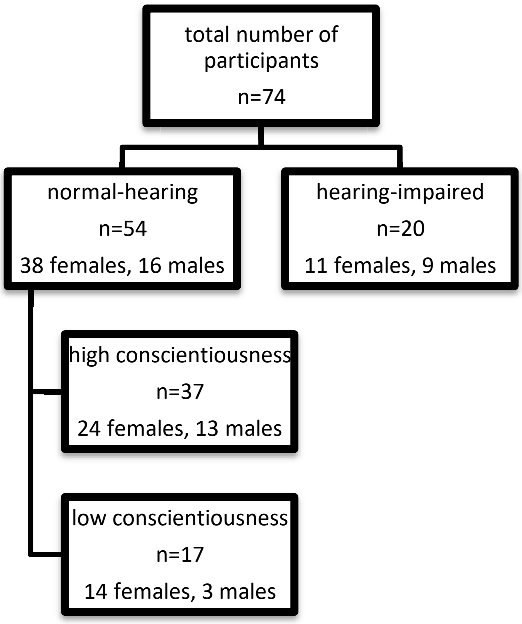


***Figure 6.5*** *Demographics of test sample in study 5*

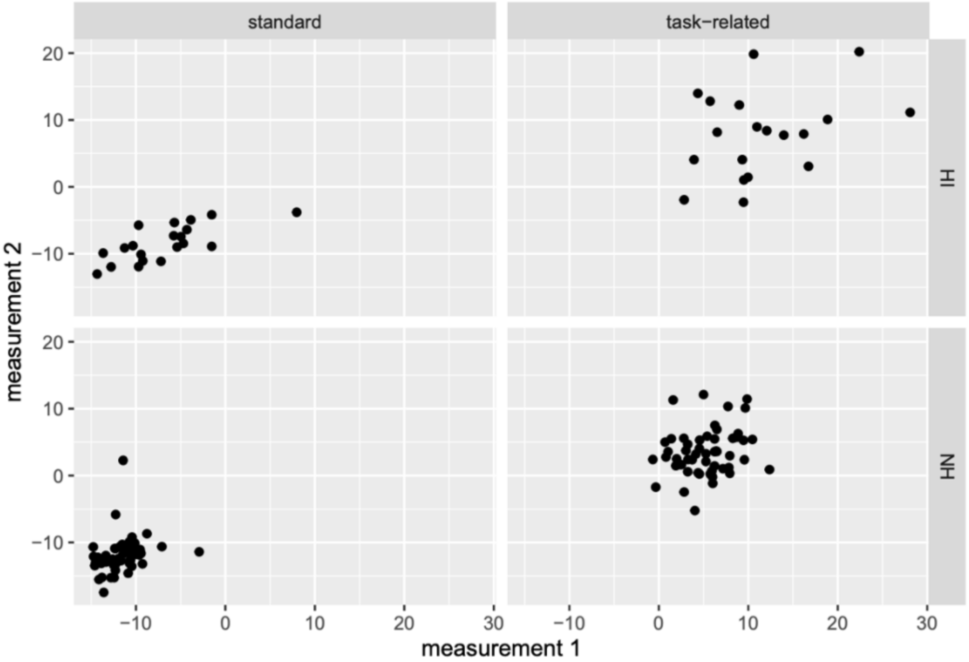


***Figure 6.6*** *Scatterplots showing the distributions of SRT scores for each condition (standard and task-related) across two measurements in NH and HI groups*

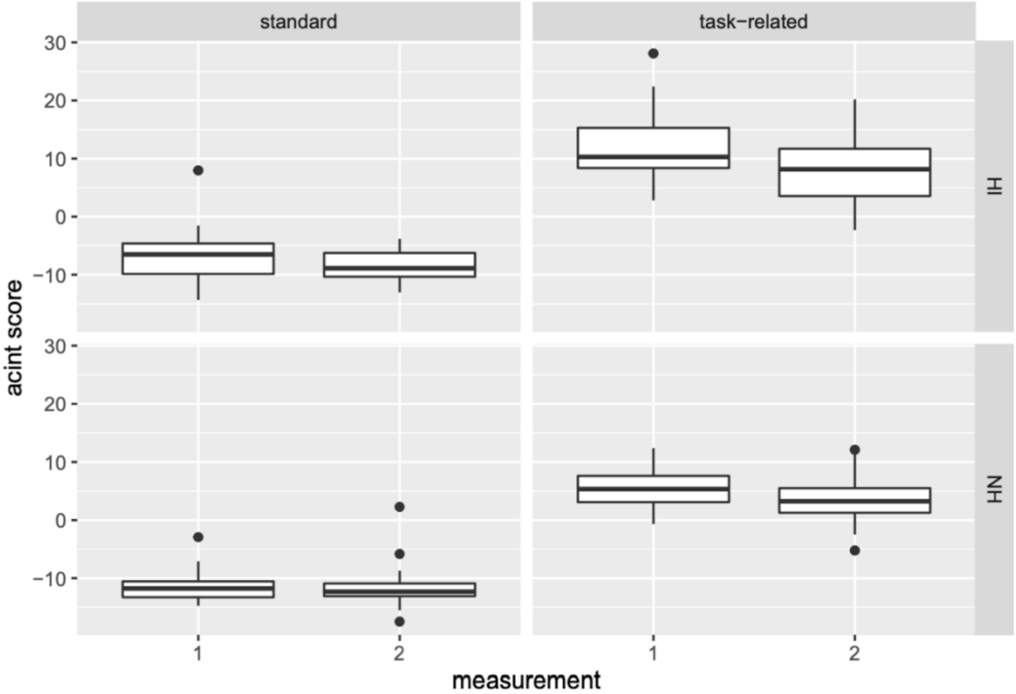


***Figure 6.7*** *Boxplots showing the distributions of SRT scores for both conditions and measurements in normal-hearing and hearing-impaired groups*

The spread between measurements is considered. Table 6.2 shows ratios of the standard deviations for measurement 1 and measurement 2, for each combination of hearing status and condition. For HI individuals in the standard condition, the standard deviation is much larger in the first measurement than in the second, almost by a factor of 2. Some of this difference is attributable to a single outlying observation, visible in the box plot in Figure 6.7. In other combinations, the standard deviations are more closely comparable.

***Table 6.2** Ratio of standard deviations (SD) for measurements 1 and 2 in both conditions*

| Status | Condition | SD1 | SD2 | Ratio |
|---|---|---|---|---|
| HI | task-related | 6.4 | 6.4 | 1.0 |
| HI | standard | 5.1 | 2.7 | 1.9 |
| NH | task-related | 3.0 | 3.5 | 0.9 |
| NH | standard | 2.1 | 2.7 | 0.8 |

### 6.6.1 Exploring test performance in normal-hearing individuals

For the normal hearing group, in addition to exploring the effects of condition and measurements, the between-subject effect of conscientiousness on task performance was assessed. A three-way mixed ANOVA was run to assess the within subject effect of conditions and measurements and the between subject effect of conscientiousness.

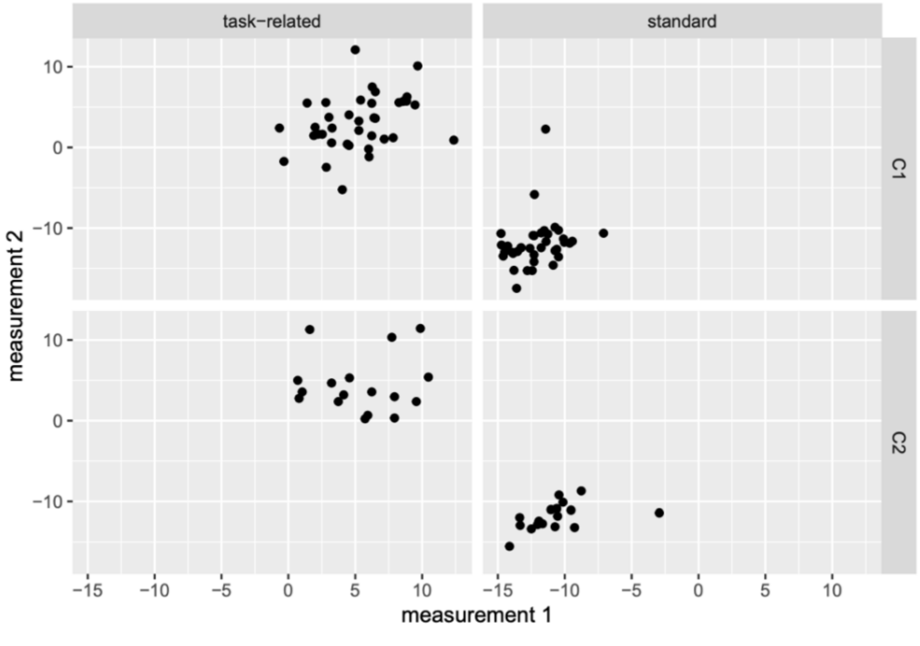


***Figure 6.8** Scatterplots showing the distributions of SRT scores for both conditions (standard and task-related) and measurements in high (C1) and low (C2) conscientiousness NH groups.*

There are four outliers in the high conscientiousness group, and three in the low conscientiousness group as assessed by inspection of the boxplots in Figure 6.8 for values greater than 1.5 box-lengths from the edge of the box. The outliers for three of the subjects were due to sudden bursts of noise during the testing session that probably affected performance. A decision was made to keep all outliers.

Violations of normality were found in the standard condition but given the good sample size and the robustness of the ANOVA, the analysis was done. Data are mean ± standard deviation, unless otherwise stated. There was homogeneity of variances as assessed by Levene's test for equality of variances (*p*>0.05).

There was no statistically significant three-way interaction between conditions, measurements and conscientiousness level, *F* (1,52) = 2.763, *p* = 0.10.

There was no statistically significant interaction between level of conscientiousness and test condition *F* (1,52) = 0.01, *p* = 0.93 or between level of conscientiousness and measurements *F* (1,52) =0.004, *p* = 0.95.

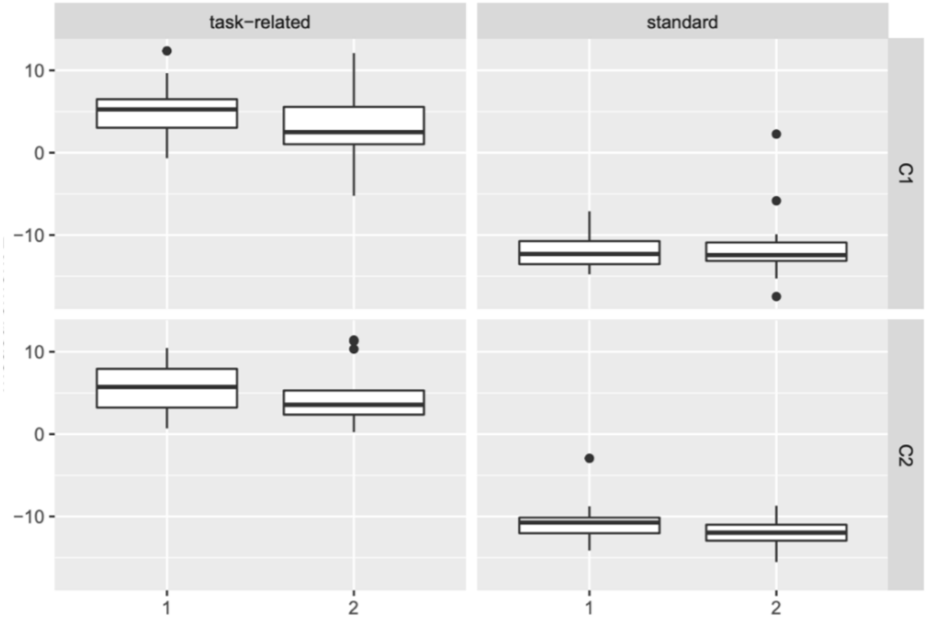


***Figure 6.9*** *Boxplots showing the distributions of SRT scores for each condition (standard and task-related) and repeat in high (C1) and low (C2) conscientiousness groups*

There was a statistically significant main effect of conditions *F* (1,52) =1816.84, *p* <0.001. There was also a statistically significant main effect of measurements, *F* (1,52) =8.29, *p* = 0.01. Simple pairwise comparisons were run between the different measurements and conditions. Bonferroni corrections were made with comparisons within each simple main effect considered a family of comparisons. Effect sizes (Cohen's d) are  given by dividing the mean difference in the measurements by the standard deviation of the difference (Lakens, 2013).

***Table 6.3*** *Factors (conscientiousness level, measurement, and test condition) considered in the three-way ANOVA including results of dependant variable (mean SRT) at each level*

| **Conscientiousness level** | | | | | |
|---|---|---|---|---|---|
| **High** | | | **Low** | | |
| **Measurement** | **Standard** | **Task-related** | **Measurement** | **Standard** | **Task-related** |
| **1** | -12.08 ± 1.74 | 5.24 ± 2.88 | **1** | -10.76± 2.52 | 5.37 ± 3.28 |
| **2** | -11.91 ± 3.11 | 3.11 ± 3.44 | **2** | -11.89± 1.68 | 4.44 ± 3.52 |

There is a statistically significant mean difference, -2.13 (95% CI, -3.45 to -0.81) dB SNR, *p*=0.002 between measurements of the task related condition in highly conscientious individuals.

In the highly conscientious group, contrasting standard and task-related conditions, a statistically significant mean increase is observed, of 17.31 (95% CI, 16.22 to 18.42) and 15.02 (95% CI, 13.71 to 16.32) dB SNR in the first and second measurements respectively.

There are no significant differences between the measurements of the standard condition or the measurements of the task-related condition in the low conscientiousness group. Comparing between the standard- and task- related condition, we found a statistically significant mean increase of 16.13 (95% CI, 14.5 to 17.75) and 16.33 (95% CI, 14.41 to 18.26) dB SNR in the first and second measurements respectively.

### 6.6.2 Exploring test performance in hearing impaired individuals

A paired samples t-test was used to determine whether there was a statistically significant change in performance on the test in different conditions across measurements in the HI sample. One outlier was detected in the first measurement of the task-related condition as illustrated in the boxplots in Figure 6.7. Inspection of the value did not reveal it to be extreme and no other reason was apparent to exclude it, so it was kept in the analysis. The differences in SRT scores in conditions were normally distributed as assessed by Shapiro-Wilk's test (all conditions had *p*>0.05). Data are mean ± standard deviation, unless otherwise stated.

Participants had higher thresholds on the task-related condition in both measurements. There was a decrease in the SRT score in second measurement of the standard condition of 1.55 (95%CI, -0.16 to 3.27) dB SNR compared to the first, but it was not significant, *t*(19) =1.89, p=0.07. There was also a decrease in the SRT score in the second measurements of task-related condition of 3.67 (95%CI, 0.08 to 7.26) dB SNR.

***Table 6.4*** *Mean SRTs of the hearing-impaired sample across conditions and measurements*

| **Hearing impaired individuals** | | |
|---|---|---|
| **Measurement** | **Condition** | |
| | **Standard** | **Task-related** |
| 1 | -6.87 ± 5.14 | 11.61 ± 6.53 |
| 2 | -8.43 ± 2.73 | 7.94 ± 6.38 |

Performance in the second measurements shows a significant decrease in SRT compared to the first in the task related condition only, *t*(18) =2.15, p=0.02, d=0.49. There is an increase of 18.48 and 16.37 dB SNR between the standard and task-related condition in the first and second measurements respectively.

### 6.6.3 Exploring the predictive ability of the generic test

To explore whether a generic test condition can predict performance on a task-related test condition, the presence of a relationship between the performance and variation in the standard and task-related condition must be determined. This will be expressed using a linear regression, as it will determine whether there is a statistically significant relationship between the two variables. It will also determine how much variation in the dependant variable (scores on the task-related condition) can be explained by the independent variable (scores from the standard condition). If a relationship is found between performance on the tests, direction and strength will be determined, and performance prediction on the task-related test from the standard test scores will be ascertained.

To determine a linear relationship, a scatterplot of the standard test scores against the task-related test scores was generated. Two data points were outliers due to disruptions during the testing session and were removed so as not to skew the regression line. The scatterplot in Figure 6.10 illustrates the plotted scores for the NH sample, with the outliers removed. A linear relationship between the test conditions can be assumed from looking at the scatterplots in both conditions.

Scores on the standard test condition did not significantly predict SRT scores on the task-related condition, adjusted $R^2$ = 0.013, $F$ (1, 50) = 1.69, $p$ > 0.05.

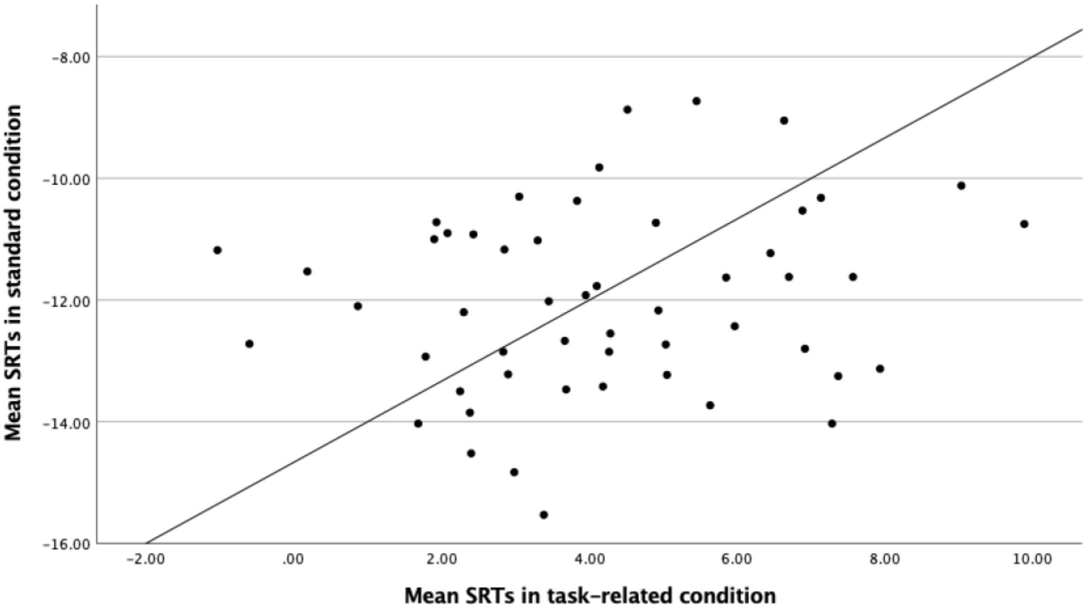


***Figure 6.10*** *Scatterplot of standard condition scores against task-related scores in normal- hearing individuals*

### 6.6.4 Exploring the test's discriminative ability in standard and challenging conditions

To assess the ACINTs ability to discriminate NH individuals from those with mild SNHL as diagnosed with PTA, a receiver operating characteristic (ROC) curve was generated.

A test's diagnostic capability is measured by its sensitivity and specificity. Sensitivity and specificity are both expressions of the accuracy of a test with binary outcomes (positive or negative). Sensitivity is the probability of correctly identifying an individual with the condition (true positive). Specificity is the probability of correctly identifying an individual without the condition (true negative) (Mathias,2010). These measures rely on a single defined cut-off point. As is the case with most newly developed tests, assigning cut-off points is challenging. A ROC curve facilitates evaluating the test across the range of possible sensitivities and specificities. It does so by non-parametrically plotting the sensitivity over (1-specificity) i.e., the false positive rate.

The ROC curve consists of three elements, the curve plotted by the sensitivity over (1-specificity) points, a 45-degree line through coordinates (0,0) and (1,1) representing chance and the area under the curve (AUC) representing the accuracy of the test. The AUC is a value between 0-1 where 0 is a completely inaccurate test and 1 is a completely accurate test. A value of 0.5 represents a ROC curve falling on the diagonal chance line therefore expressing no discriminative ability. A value between 0.8-0.9 is acceptable and a value above 0.9 is excellent (Mandrekar, 2010).

For this analysis, the repeats in both conditions were averaged giving us the mean score for each test condition. This analysis was done for conceptual reasons with the acknowledgement that it is not ideal to compare SiN performance to PTA, which is why the averaged SRTs were used for ease of analysis and interpretation. Additionally, the focus is on the standard condition since the task-related condition is under-developed and was used for preliminary exploratory purposes. Since there was no difference between the repeats of the average condition, it was feasible to average the results. A ROC curve illustrated in Figure 6.11 was generated to assess the diagnostic performance of the test conditions. The blue line represents the diagnostic performance on the standard test condition and the green line represents performance on the task-related test condition. The AUC is equal to 0.85, 95% CI (.76,.98) p <0.001 for the standard condition and 0.83, 95% CI (.72,.96), p<0.001 for the task-related condition. They are both significantly different from 0.5 indicating that both test conditions could distinguish between the NH and HI samples. If the highest points (the point at which both coordinates are furthest from the chance line) in both curves are assumed as the cut-off points, this would give the following cut-off points:

- **standard condition:** -10.86 dB SNR with coordinates (0.26,0.84) on the ROC curve.
- **task-related condition:** 5.16 dB SNR with coordinates (0.33,0.84) on the ROC curve.

The calculations for the accuracy measures reported in Table 6.5 were made based on calculating the true and false positive and negative rates as identified by the chosen cut-off points.

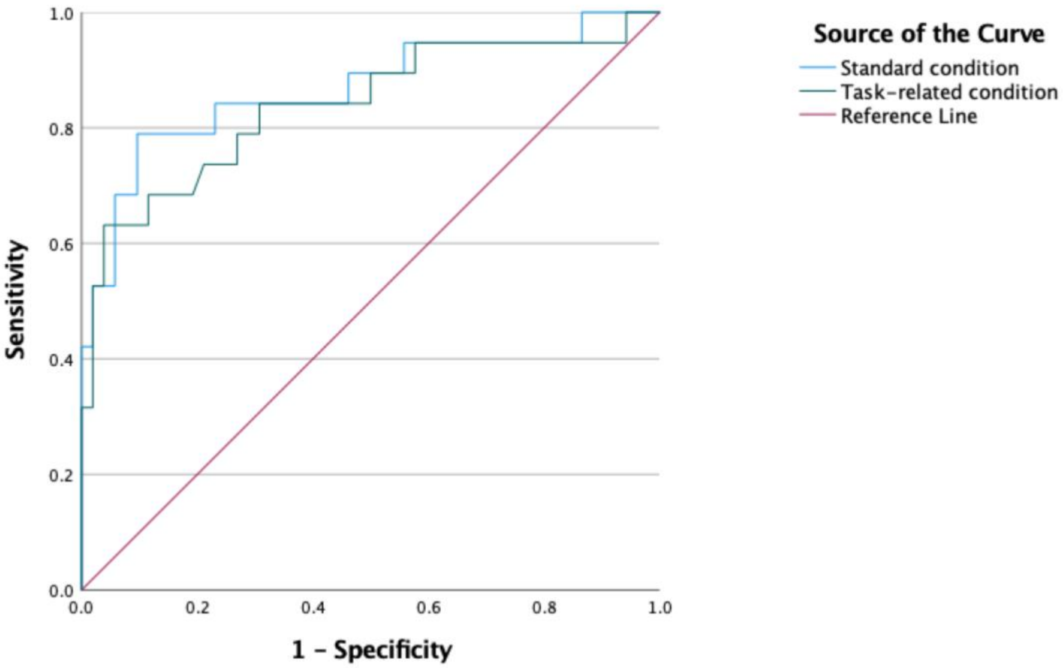


***Figure 6.11*** *ROC curves for the standard and task-related conditions*

***Table 6.5*** *Accuracy measures for each test condition based on the chosen cut-off points from the ROC curve*

| **Test condition** | **Standard** | **Task-related** |
|---|---|---|
| accuracy | 0.77 | 0.71 |
| error rate | 0.23 | 0.29 |
| sensitivity | 0.85 | 0.84 |
| specificity | 0.74 | 0.67 |

Table 6.6 reports the positive and negative predictive values based on the chosen cut-off points. The positive predictive value is the percentage of true positive patients testing positive. it is calculated using the following equation from the values in Table 6.6.

$$PPV: = a / a+b$$

Where *a* is true positive, and *b* is false positive

The negative predictive value is the percentage of true negative patients testing negative it is calculated using the following equation from the values in Table 6.6.

$$NPV: = d / c+d$$

where *d* is true negative, and *c* is false negative

***Table 6.6*** *Predictive values based on the chosen cut-off points in the ROC curve*

| **Standard condition** | **Predicted positive** | **Predicted negative** |
|---|---|---|
| **True positive** | (a) 17 | (c) 3 |
| **True negative** | (b) 14 | (d) 40 |
| | | |
| **Task-related condition** | **Predicted positive** | **Predicted negative** |
| **True positive** | 16 | 3 |
| **True negative** | 18 | 36 |

Based on our values, the test has 54.8% PPV and 92% NPV in the standard condition.

### 6.6.5 Assessing test repeatability in NH individuals on the standard condition

The goal in development of any measure is to ensure it is valid and reliable, meaning that it measures what is intended to measure and does so consistently with minimal measurement error (Bartlett and Frost, 2008). Having a repeatable test is very important to ensure precision of the measurement. It is also an important step of assessing test reliability.

Repeatability, the amount of within-subject variation in repeated measures, not subject to temporal or test condition changes, is assessed by looking at the within-subject standard deviation in the measurements (Vieira and Corrente, 2011). It is normal to have some intra- and inter-individual variation in the results when taking the test more than once, otherwise known as measurement variability (Watson and Petrie, 2010). What is important is to establish the amount of acceptable measurement variation in repeatable results. This is done by:

- ensuring the absence of a systematic difference, assessed by a t- test or ANOVA.
- ensuring agreement of the measurements within defined limits of agreement assessed by Bland-Altman plots

Bland and Altman (1986) proposed quantifying the agreement between two measures from their mean differences. This quantification is then displayed in a Bland-Altman (BA) scatterplot which plots the mean of both measures on the x-axis against the differences between each measure on the y-axis and defines limits of agreement based on the standard deviations of the differences. This method only identifies the limits of agreement which are within ±2 standard deviations for 95% of the data. The resultant values can then be assessed based on prior knowledge of acceptable test values (Giavarina, 2015). A Bland- Altman plot is constructed in Figure 6.12. Shown is the difference in the two measurements, plotted against their average. The dashed red line indicates the mean difference. The blue dashed lines are lower and upper limits of agreement at the mean difference ±2 standard deviations. Assuming the difference to be normally distributed, ~ 95% of observations should lie within the two lines. For these data, 94 % lie within the two lines; there are no more observations outside the lines than would be expected by chance. Seventy-seven percent of the data lie within ±1 SD equivalent to ±2.5 dB SNR from the mean.

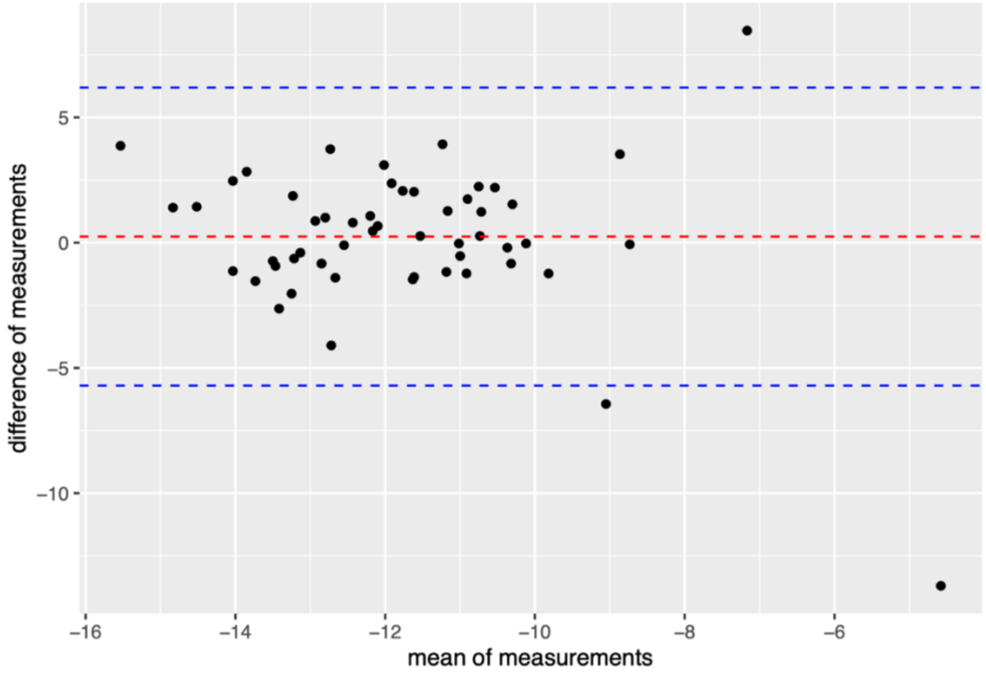


***Figure 6.12*** *Bland-Altman plot illustrating the repeatability of the test in the NH group on the standard condition*

### 6.6.6 Exploring the relationship between SiN test scores and PTA

The relationship between SRT's in noise and PTA scores is explored through correlations. The Pearson correlation coefficient, also known as the Pearson product-moment correlation coefficient, is a standardized measure of bivariate correlation yielding a value between -1 and 1, indicating a negative or positive correlation respectively. Squaring the coefficient can give further information about the amount of variability shared between the two assessed variables.

A Pearson's correlation was run to assess the relationship between SiN test scores and a range of PTA configuration scores in both conditions and test samples. The range of configurations was chosen based on the variation in the associations found in the literature (Smoorenburg, 1992; Phatak *et al.,* 2019). Results of 46 normal-hearing and 20 hearing-impaired participants were correlated. The PTA done on the NH sample in the schools was for screening purposes using a portable machine. The machine did not save or print the results. In one school, all the students were screened in the morning prior to administering the test and called back individually to take the test. Since the primary goal of the PTA was screening and recruitment of a larger number of participants was prioritized, their full PTAs were not documented. Correlations were done between SRT's and the following PTA score averages: 0.5, 1 and 2 kHz; 1, 2 and 4 kHz; 2 and 4 kHz; 2, 4 and 8 kHz; and 4 and 8 kHz. The rationale for choosing a wide range of configurations was the differences across studies correlating SRTs and PTA results. Preliminary analyses showed the relationship to be linear with both variables normally distributed, as assessed by Shapiro-Wilk's test (*p* > .05). Correlation between SRTs and each PTA configuration were done separately to type I errors.

Correlations are demonstrated as pair plots shown in Figure 6.13 and Figure 6.14 for NH and HI groups respectively. Pair plots are simple matrices used to visualize the relationship between each variable in the data set (Emerson *et al.*, 2013), in this case the PTA configurations and SiN test conditions.

Pair plots show the correlation between SRTs and PTA, by frequency. For SRTs, the average of the two measurements in each condition is used. In the grid of plots below, the diagonal gives a smoothed histogram (density plot) showing the marginal distribution of each variable. The pair plot shows the joint distribution of two variables. Reflected for example in Figure 6.13, in the first row, third column, is the highest correlation between the average PTA 0.5-1-2 score and the average standard condition SiN score. This correlation is 0.59, explaining 32% of the variance in SRT's. The three stars denote statistical significance at the 0.1% level, two stars at the 1% level and one star at the 5% level.

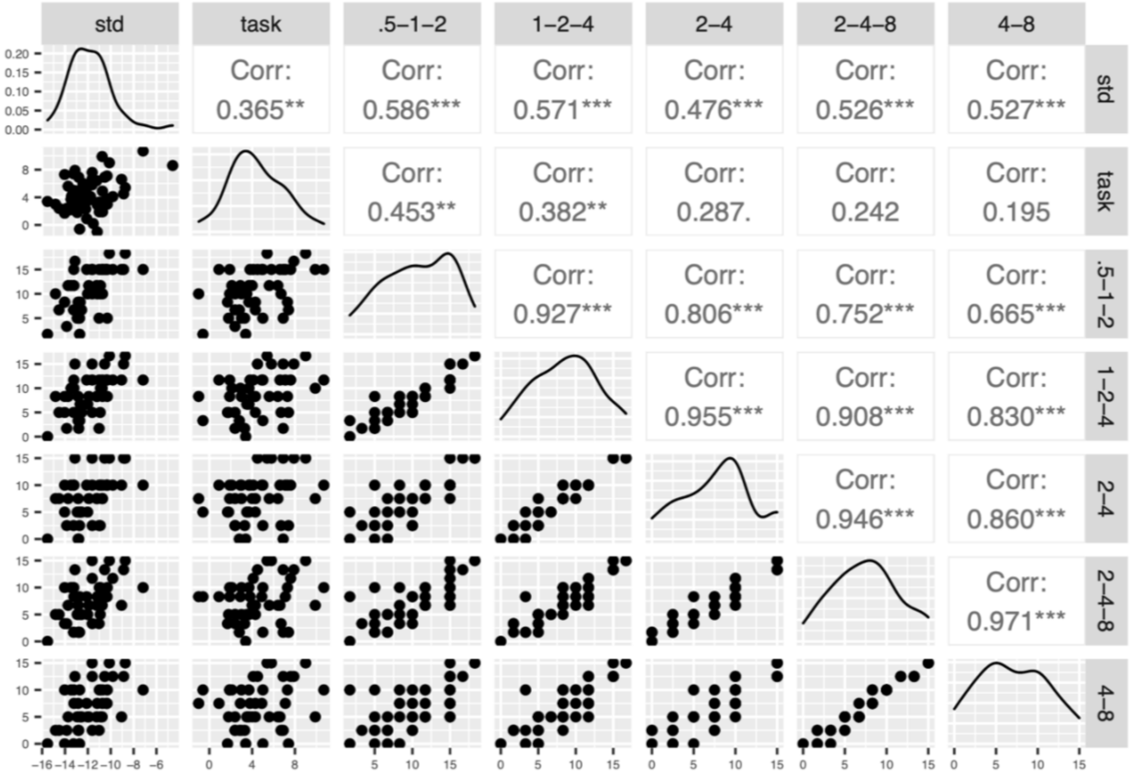


***Figure 6.13*** *Pair plots illustrating the correlation between SRTs and PTA averages in both test conditions in the NH sample. The diagonal line represents the density plots for each variable*

In the standard condition, for NH individuals, there is a strong positive correlation between SRTs and all PTA configurations, apart from 2-4 kHz which showed a moderate correlation. Moderate correlations were found in the task-related condition only in PTA configurations 0.5-2 and 1-4 kHz.

There are only HI 20 subjects, so what seem to be large correlations are not statistically significant. A correlation was found between the standard test condition and PTA average scores for 2 and 4 kHz *r*(18) = .52, *p* = .02 (figure 6.12). No correlations were found between the task-related test condition and PTA in the HI sample.

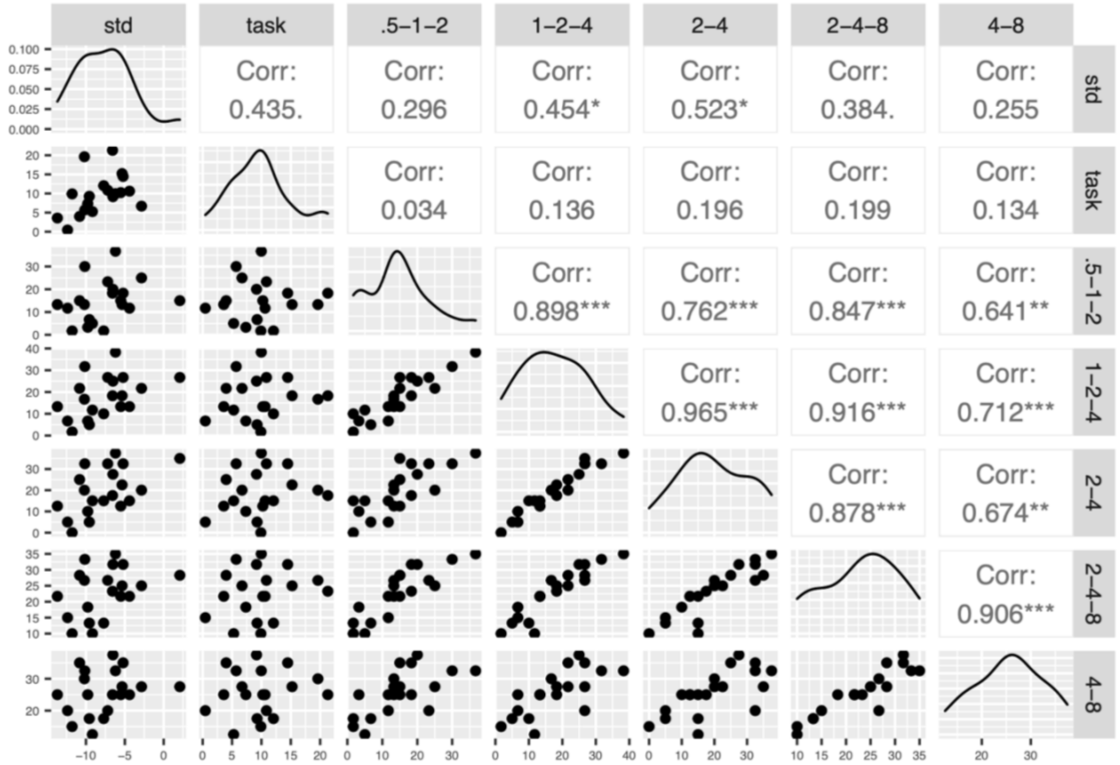


***Figure 6.14*** *Pair plots illustrating the correlation between SRTs and PTA averages in both test conditions in the HI sample. The diagonal line represents the density plots for each variable.*

## 6.7 Discussion

This study aimed to evaluate a newly developed SiN test for AFFD testing purposes and explore the performance of a relevant sample and factors that could affect test performance. The results of the following aims will be discussed separately:

- Exploring performance differences on the ACINT under standard and task-related listening conditions in NH and HI individuals.
- Exploring the relationship between conscientiousness as a source of possible variation in NH individuals.
- Assessing if performance in the task-related condition can be predicted from performance in the standard condition.

- Examining the sensitivity of both test conditions at detecting mild SNHL.
- Assessing the repeatability of the standard test condition in NH individuals.
- Exploring the correlation between PTA and SiN performance and compare it to existing literature.

### 6.7.1 Performance on standard and task-related conditions and the effect of conscientiousness on variation in performance in a NH sample

In the NH group, there was a significant difference between performance on the standard and task-related condition, with an average improvement of 16 dB SNR in the standard condition compared to the more challenging task-related condition. Attenuation of the frequencies in the task-related condition contributes considerably to the decline in performance as a marked portion of relevant speech information found in the higher frequencies is degraded.

Both conditions had a background of SSSN. Several tasks carried out by the RSADF are done in proximity to a noisy generator which is similar to SSSN albeit easier, because it lacks the difficulty added by the noise being speech shaped.

The improvement in performance between measurements, representing the learning effect was present in the task related conditions but was significant only in the highly conscientious individuals. This is consistent with studies that have shown the pattern of learning to be similar across different test conditions, with more challenging conditions having a larger learning effect (Cainer, James and Rajan, 2008; Rhebergen, Versfeld and Dreschler, 2008; Schlueter et al., 2016). The method followed for overcoming the learning effect appears to be sufficient for the standard condition. Degraded listening conditions appear to require more training prior to starting the test. Perhaps the effect was not elicited in the low conscientiousness individuals due to their smaller number. It is also possible that there is a small effect for conscientiousness in the more challenging condition that was not elicited in the study due to the small number of low conscientiousness participants. However, when both low and high C groups were combined, a difference between measurements was still present (1.54 dB difference between measurements) with an effect size 0.44. This is generally considered a small effect.

In the HI sample, the learning effect was also present in the task-related condition (3.67 dB between measurements). The effect size is slightly larger, 0.49. Rhebergen, Versfeld and Dreschler (2008) attributed variation in performance to the increase in speech information as a function of SNR. In their study, they were comparing between stationary and non-stationary types of noise. However, it appears the degradation in the speech signal produces a similar effect.

Although findings are mixed in the literature, in our study, consistent with others, it appears HI individuals may require more training than NH to reach a stable performance level (Hagerman and Kinnefors, 1995; Burk *et al.*, 2006). Variation in performance across individuals is also larger in the HI sample and improvement in the second measurement is larger in HI group in both conditions. There was a mean difference in performance between NH and HI individuals of 4.01 dB SNR in the standard condition, and 5.23 dB SNR in the task-related condition. Bronkhorst and Plomp (1992) found the mean difference in performance between NH and HI listeners to range from 4.2-10 dB across different babble maskers. The decrease in performance from the telephone effect is larger for HI individuals.

Regarding the effect of conscientiousness, there is a small, yet consistent effect documented mostly in organizational literature interested in the task component of job performance. However, no significant effect was found in this study. This could be due to one or more of the following reasons:

1. The conscientiousness -task performance relationship may be subject to temporal effects (Meyer, Dalal and Hermida, 2010) and time pressure. Perhaps if the participants were subjected to time pressure, a relationship, if present, would be made more obvious.
2. The nature of the conscientiousness -task performance relationship may be very context specific, and not apparent in all tasks and situations (Meyer, Dalal and Hermida, 2010; Perry et al., 2010; Moldzio et al., 2021; Wang, Liao and Burns, 2021). Perhaps, had it been studied on a military sample aware that the task was duty-related, an effect would have been elicited. Perhaps also, had a context specific measure of conscientiousness been used, as suggested by Moldzio et al., the relationship may have been more apparent.
3. The relationship may be facet specific and not related to the global trait (Dudley et al., 2006; Darr and Kelloway, 2016; Moldzio et al., 2021) which is measured by the NEO-FFI.
4. The inequal sample sizes between the low and high conscientiousness groups and small number of low conscientiousness participants may have contributed, which is not surprising, given the nature of low conscientiousness individuals not being keen on volunteering in studies as their more highly conscientious counterparts. As discussed in 6.4.4, the required number of participants in each group for an experimental power of 0.9 was 20, and the participating low conscientiousness individuals were 17.

5. The direct relationship between conscientiousness and SiN performance may not exist because SiN recognition is so fundamental to human communication, that it is resolved at the brain stem level and not affected by conscientiousness. However, the relationship could be mediated by factors such as attention and motivation, which were not measured directly in the study.

### 6.7.2 The relationship between standard and task-related conditions

If a broader test can predict performance of a more specific test, it would make it more applicable and favourable. To assess the predictive ability of the test in the standard condition, we looked at the NH sample only, as the sample size of the HI group is too small. There is a relationship between the two conditions as seen in the moderate positive correlation between them, but no significant predictive relationship was found from the linear regression. Previous studies have found performance in SSSN to be predictive of performance in real-world noise (Wong, Ng and Soli, 2012) or amplitude modulated noise (Schoof and Rosen, 2014). This may not have been found in this study because degradation of the signal has a different effect than degradation in the noise background. Moderate correlations have been found between performance in SSSN noise and noise-vocoded speech ($r=0.56$, $p < 0.05$)(Paulus, Hazan and Adank, 2020). Perhaps an important factor is the difference of the listening range in both conditions where the task related condition is focused on mid frequencies and is quite different from the standard condition.

### 6.7.3 The discriminative ability of the test and the relationship between SiN and PTA scores in standard and task-related conditions

The obtained AUC values, 0.85, $p <0.001$ for the standard test condition and 0.83, $p<0.001$ for the task-related test condition, representing test accuracy are comparable to accuracy scores of other SiN tests in the literature, namely the US national hearing test (NHT) and the words in noise (WIN) test, having AUC scores of 0.82 and 0.89 respectively (Williams-Sanchez *et al.*, 2014). Both tests measured their accuracy against PTA, on the basis that it is the traditional gold standard in diagnostic audiology. They also compared the NHT to WIN and while comparison of like measures is better, it unfortunately was not possible in our study.

Perhaps the accuracy and the subsequent calculations in our study may have been higher had the criteria applied for the definition of hearing loss not been so stringent. Some studies accept ≤ 25 dB as normal hearing (Sheikh Rashid and Dreschler, 2018; Essawy *et al.*, 2019; Paulus, Hazan and Adank, 2020; Elrifaey *et al.*, 2021), whereas the three individuals who were false negative had a dip at 8 kHz frequency only at 25dB, a frequency not measured in all SiN test development and validation studies (Bentler, 2000; Killion *et al.*, 2004; Xi *et al.*, 2012; Houben *et al.*, 2014; Canzi *et* *al.*, 2016; Sheikh Rashid and Dreschler, 2018). Then again, others consider classifying thresholds of 20 dB as normal questionable (Pienkowski, 2017). Several recent studies have demonstrated the association of high frequency speech information (Trine and Monson, 2020; Polspoel *et al.*, 2022) and extended high frequency (EHF) PTA with speech in noise recognition (Motlagh Zadeh *et al.*, 2019; Mishra, Saxena and Rodrigo, 2022). If EHFs have proven to play a role in SiN recognition, then the minimum acceptable range of PTA testing when researching SiN should be up to 8 kHz.

The issue arising with reference standards in general and ours particularly is measurement error. This is an occurring problem when the reference standard test is not the ideal comparator, or a true reference standard is not available (Zou, O'Malley and Mauri, 2007). Ideally, the developed test should be assessed against a reference SNR defined by the task and noise environment analysis. This was done by a few institutions including the California Corrections Standards Authority (CSA) and the Ontario Ministry of Community Safety and Correctional Services, where noise recordings of identified noise environments were obtained, participants undertook the HINT and resulting data points were used to produce speech intelligibility indexes (SII). From these calculations, SII-intelligibility transfer functions were obtained and ESII values in each noise environment were determined (Soli, Amano-Kusumoto, *et al.*, 2018).

An important distinction to make is that between measures of the test; sensitivity and specificity, and measures of the practical usefulness of the test on people; PPV and NPV. Positive predictive value expresses the probability of individuals testing positive truly having the condition. Negative predictive value expresses the probability of individuals testing negative truly free of the condition of interest (Trevethan, 2017).

Test measures provide information of the test performance relative to the chosen reference; in our case the PTA, whereas PPV and NPV reflect the practicality of the test in a clinical setting. The test in the standard condition has 54.8% PPV and 92% NPV. An important consideration when evaluating PPVs and NPVs is the disease prevalence. The PPV calculated values will be higher in diseases with higher prevalence and this affects interpretation. It is important that the study dataset mirror the prevalence of the disease condition in the population. Considering this, the prevalence of HI in our study sample is 37%. While the NH group is representative of entry-level military candidates, the HI group is not. This makes it difficult to interpret these values. It is also difficult to accurately estimate the prevalence of hearing loss in the military because the data provided by military health sectors is insufficient. A study assessing NIHL in Saudi military personnel, screened 1879 service members attending tertiary care centres with a mean service period of 10.26 ±8.06 years using a noise exposure survey and PTA, defining HL at >25 dB at frequencies 0.25-8 kHz. Based on the screening criteria, 10% were identified to have NIHL, with 4.7% being from the air defence (Al-Saif, 2014).

The interpretation of PPVs and NPVs also depends on the purpose of the test. A test with higher PPV indicates minimal false positive outcomes. This is desirable if a positive result is a source of anxiety or harm, financial burden to the employer, health-care system, or the test-taker. A test with high NPV is desired in cases where the tested condition needs to be identified early to prevent progression, or in serious or contagious conditions. The chosen cut-off point used to calculate all the accuracy measures was chosen for demonstrative purposes to be the point with maximum sensitivity and specificity. In actuality, the test purpose must be carefully considered, and performance criterions must be established for the cut-off point to be based upon. Returning to the purpose of our test, a high PPV and NPV is clearly favourable as a false positive could be detrimental to the employee's livelihood. No developed SiN tests in the literature were found to report PPV or NPV values, even those developed for AFFD purposes.

While the sensitivity and specificity results are initially reassuring, they are not to be misinterpreted. The discriminative ability of the test merely informs us of the test's sensitivity to the characteristics of the assessed population. It does not indicate the quality of the test as a diagnostic tool. This has yet to be done and requires a larger scale validation and reliability study done across suitable time intervals by more than one administrator and on more populations. Regarding the results obtained in the ROC curve, another perspective is that the subjects identified as false positive can be viewed as the individuals who don't have the necessary SiN recognition capabilities and those identified as false negative can be seen as the individuals have necessary hearing abilities for conditions in the test.

This is all in comparison with PTA as a reference. Circling back to the important point that PTA is not the ideal reference for AFFD, there is a need for comparison to a well validated similar SiN test, or obtainment of the exact SNR employees need to be able to discriminate at to be used as a reference standard criterion, either by testing with fixed SNRs or using the chosen point as the cut-off point.

### 6.7.4 Test repeatability

Test repeatability assessed by Bland-Altman plots illustrated that 94% of the data fall within the limits of agreement with 77% of the data lying within ±1 SD equivalent to ±2.5 dB SNR from the mean. Had the outliers been removed, more than 95% of the data would have fallen within the limits of agreement. Resulting repeatability is comparable to the values obtained from the simulation study (study 3) exploring the range of acceptable homogeneity. It is also comparable to other developed SiN tests in the literature (Nilsson, Soli and Sullivan, 1994; Ozimek *et al.*, 2009).

There are two types of errors that must be estimated in any test, systematic error, and random error. Systematic error is the difference between the measured parameter and the parameter's actual value, otherwise referred to as bias (Wells and Sireci, 2020). Accurate tests are low in systematic error. This was assessed in our test by the ANOVA. Random error is the difference in repeated measurements under the same conditions. This informs us of the precision of the test.

The test has proven to be low in random error. This is reassuring as it is a sign of consistency of test performance within and between subjects. This is the first step to ensuring test reliability. Building on these results, the test should now be assessed for test-retest reliability which involves repeating the measurement after a certain amount of time has passed. It should also be repeated on the target population.

### 6.7.5 The relationship between PTA and SiN scores

The literature regarding the relationship between PTA and SRTs is quite conflicting. Some studies found a strong relationship between them, others found a relationship only with higher frequencies and some studies found no relationship. In the standard condition, for NH individuals, there was a strong positive correlation between SRTs and all PTA configurations, apart from PTA configuration 2-4 kHz which showed a moderate correlation. Moderate correlations were found in the task-related condition only in PTA configurations 0.5-2 and 1-4 kHz.

As for the HI individuals, in the standard condition a strong correlation was found for PTA averages from 2-4 kHz and no correlations were found between PTA and SRT in the task-related condition. This is not surprising because the speech spectrum looks like an inverted banana with a roll off at higher frequencies. Since the audibility and distortion components of hearing correlate more in these regions, listeners with sloping hearing losses at higher frequencies may have an increased correlation between PTA and SiN results. This correlation with PTA is also more pronounced in low level signal and adaptive tests such as ours as the audibility component also plays a large role.  This effect is less pronounced in higher level suprathreshold stimuli presented at a fixed SNR as this falls more in the distortion domain (Brungart *et al.*, 2022).

In most studies regarding AFFD, there is a persistence that PTA does not capture the performance of individuals listening to speech in a noisy environment. This is evident in our results where correlations were found between both measures, but a large portion of the variability in the SiN score was still unexplained, reinforcing the argument that other factors affect SiN. Other factors which contribute to AFFD, such as experience and cognitive abilities are important in SiN performance in an AFFD context. An individual may for example, have mild hearing loss on a PTA but be able to perform their job, and listen to commands in a noisy environment. This fortifies our advocacy for inclusion of SiN tests in AFFD assessments.

## 6.8 Conclusions

This study explored the developed SiN test; the ACINT, as a suitable tool for assessing AFFD in a military population. Unfortunately, developing an AFFD standard for our military population was beyond the scope of this PhD. When evaluating functional assessments of AFFD, considerations must be made to the fit of the conceptual model of the test and any required cultural adaptations, its validity and reliability, sensitivity and predictive value, ease of interpretation and administrative burden (National Academies of Sciences, Engineering, and Medicine, 2019). The developed test has proven to have validity and sensitivity initially. To be reliable and fit conceptually it must be tested on the target population. It is a low burden test that can also be adapted into a more accessible format (tablet) for further testing.

The ACINT proved to be a repeatable test, suitable for further assessment. Future work should build upon these results to assess the reliability of the test and determine required performance criterion on the test in different noise backgrounds. The test in the task-related format was done for exploratory purposes and is quite preliminary. It does not adequately represent the task executed by the RSADF in training and requires further evaluation. However, it did give insight regarding the effect of degradation in the signal. A possible option with degraded signals is to test using a lengthier training protocol to determine if the observed additional variability is an effect of training.

# Chapter 7 Summary, conclusions and future research

## 7.1 Summary

This thesis set out to develop a SiN test designed for military populations and explore it as an AFFD assessment tool. This required a certain amount of cooperation from the military and detailed task-analysis of the chosen sectors duties. Due to unforeseen circumstances, namely the COVID-19 pandemic and the subsequent ending of collaboration with the RSADF from the military's side, this high level of co-operation could not be seen through and instead the body of research focused on developing a general SiN test adapted from the CRM. The developed test had general military characteristics in terms of the nature of the communicated sentence structure, but similarly to the CRM can be used for non-military purposes.

The first phase of the PhD focused on test development. This was done by developing speech material using the MoCS and then optimizing it using an interleaved adaptive procedure. A study using MCSs was also conducted to explore the acceptable amount of variation in homogenous speech material. This phase was advantageous in that the development and optimization process was very thorough as it was done using two methods. Although this was time-consuming, it was also needed given the limited availability of language and dialect specific speech material. The shortcoming of this phase was that after optimization of the speech material, a complete validation study of the speech material in the final test format was not carried out and the speech material was adapted into the test format based on pilot results of 13 participants. A validation study had been intended, but due to time and COVID-19 constraints, it could not be seen through to completion. However, in retrospect, the results of the final study conducted on 53 participants, the simulations run 10,000 times and the pilot on the optimized speech material all gave similar results, indicating that the developed speech material was indeed valid and suitable for use.

The second phase of the PhD focused on exploring the developed speech material in an adaptive sentence test named the ACINT on a sample representative of entry-level military recruits. Prior to the final study, personality trait conscientiousness was chosen as factor to be explored in relation to SiN test performance. Although at this stage it had become clear that extensive collaboration would no longer continue with the military, they were still willing to participate to a lesser degree. Since the focus of the PhD was on AFFD in a military population, it was decided to explore differences in conscientiousness variation between military and civilian samples to assess the feasibility of using conscientiousness as an indicator of performance variation in SiN tests.

Results showed a small difference in variation of trait between sectors with the military sample exhibiting larger variation. This study was advantageous in its large sample size. It was however, limited by the unavailability of a valid translated version of the adolescent NEO-FFI which is more suitable for individuals between the ages of 12-20 years.

The final study focused on exploring the ACINT in its standard format of clean speech in SSSN and a more task representative condition where the speech was degraded to sound like telephone speech in SSSN. This condition was chosen to represent one of the tasks performed by the RSADF during training in the academy. A detailed task analysis was not performed, so it is unclear how frequently this task is performed, and the exact frequencies attenuation occurs at. Consequently, this condition is a crude representation at best, meant to give general insight into the effects of degraded speech in the ACINT. This study was also limited by the study sample, which was a representative sample of secondary school students and fresh graduates, meant to represent entry-level military recruits. Ideally, the representative sample could have been only students interested in pursuing a military career but given the difficulty of recruiting participants and testing in schools during COVID-19 restrictions, this was not possible. The ACINT was administered in its standard and task-related condition, measuring each condition twice to assess test repeatability. The NH sample was divided according to their level of conscientiousness into high and low conscientiousness to explore the trait's relationship with SiN test performance. No effect of conscientiousness measured by the NEO-FFI was found on test performance in both test conditions. A HI sample was also recruited to assess the test's sensitivity to mild SNHL as diagnosed by PTA. The HI sample was also limited in that it was not age-matched to the NH sample due to recruitment difficulties. The results of the final study demonstrated the repeatability of the ACINT in its standard format and its sensitivity to mild SNHL. The ACINT in its standard format is a valid repeatable SiN test suitable for further audiological research.

## 7.2 Conclusions

1- An Arabic speech corpus was carefully developed, suitable for further audiological research in different noise backgrounds.

2- Monte Carlo simulations are informative estimators of test parameters suitable when testing large samples is not possible. Using simulations, for the developed speech corpus, an SRT range within ±3 dB SNR was proven to be homogenous.

3- Insight was gained regarding trait conscientiousness in a Saudi military population.

4- The speech corpus adapted into a SiN test (ACINT), is valid and repeatable in its standard format. It is suitable for further research as a GPT test of AFFD through detailed job analysis and selection of cut-off points and noise backgrounds accordingly.

5- The global trait conscientiousness was not found to affect test performance. Studying the relationship between specific facets in more context specific test conditions should be considered.

6- Further research is required for the ACINT under degraded conditions as the degraded condition used in the final study exhibited some variability in the results. It is important to consider lengthier training protocols in degraded test conditions to determine if an element of variability is due to the effect of training.

## 7.3 Future research

Future work should build upon the results from the studies conducted in this research to assess the reliability of the ACINT. If the test is to be researched for AFFD purposes, required performance criterion on the test in different noise backgrounds should be determined depending on results of detailed task analyses.

Face and content validity of the ACINT in relation to target populations should also be assessed. This can be done by conducting focus groups and having subject matter experts (SMEs) rate the test items in terms of having characteristics relevant to the job nature.

Due to the acoustic nature of military environments and the importance of using HPDs, it is important to consider their use and effects when developing test performance criteria.

Another consideration in jobs requiring AFFD is integrating auditory training programmes whether for rehabilitative purposes for individuals who are found unfit due to their hearing or as part of the job training to improve competence. Most of the previous literature has focused on pure auditory training tasks which don't really transfer to anything outside the context of the training material. However, more recent auditory training literature has focused on combined auditory cognitive training approaches (Ferguson and Henshaw, 2015; Whitton *et al.*, 2017).

Further developing the test material to expand its uses for testing speech comprehension can also be considered. Additionally, modifying test presentation to incorporate a memory task may broaden the developed material's utility.

The MCSs studied the test parameters using one test format and a fixed number of word-options and parameters. Simulating different test formats with different step-size rules and exploring fewer or more word-options would be worthwhile.

# Appendix A Studies exploring the relationship between PTA and SIN tests

| study | sample | PTA configuration | SIN test | age | other tests | hearing status | Results |
|---|---|---|---|---|---|---|---|
| Merten et al.,2022 | 2585 | Average of 0.5, 1, 2, 4 | Göttinger Satztest | 30+ | crystallized intelligence, executive function, WM, long term memory | NH and HI | PTA strongest predictor of SIN >30 years |
| Vermiglio et al,2020 | 270 | 0.5-6 kHz | HINT | | | NH and HI | Audibility proportion affects SIN-PTA relationship |
| Holmes and Griffith,2019 | 97 | 4-8 KHz | matrix | 18-60 | auditory figure ground perception (central auditory processing) | only NH | SIN related to PTA and central grouping processes |
| Phatak et al, 2019 | 288 | 250-8 kHz | SPRINT<br>NU-6<br>Military call signs acquisition test | 18-55 | Spatial attention test<br>Functional hearing questionnaire | NH & HI | Moderate correlation between PTA and SIN. Not enough for assessing supra-threshold deficits |

Appendix A

| **study** | **sample** | **PTA configuration** | **SIN test** | **age** | **other tests** | **hearing status** | **Results** |
|---|---|---|---|---|---|---|---|
| Gieseler et al, 2017 | 438 | 0.5-4HHz | Goettingen sentence test | 60-85 | loudness scaling, cognitive tests | only HI | PTA predictive of SIN in HI hearing aid non-users |
| Stenbäck, Hällgren and Larsby, 2016 | 86 | 125-8 kHz | Hagerman SIN test | 20-75 | Cognitive tasks | NH & HI | Moderate correlations between SIN and PTA in HI |
| Schoof & Rosen, 2014 | 28 | 0.5-6 kHz | IEEE sentences | 19-29<br>60-72 | Temporal processing, cognitive tasks<br>Processing speed | NH | PTA moderately correlated to speech in babble noise |
| Vermiglio et al, 2012 | 215 | 0.25-6kHz | HINT | 17-59 | Speech in Quiet | NH to profound HFHL | Weak correlations between PTA-SIN performance |
| Barrenäs and Wikström, 2000 | 1895 | PTA-High (3, 4, and 6 kHz) and PTA-Mid (0.5, 1, and 2 kHz) | swedish PB word lists | 17-89 | Speech in Quiet | NH or SNHL | High frequency PTA predicts SIN performance |
| Smoorenburg, 1992 | 200 (400 ears) | above 3 kHz, from 1 to 3 kHz and below 1 kHz | Plomp and Mimpen test | 18-55 | Speech in Quiet | NIHL and NH working in noise | High frequency PTA predicts SIN performance |

# Appendix B Neo Five Factor Inventory-3 (NEO-FFI-3) (McCrae and Costa, 2010)

**NEO-FFI-3 consists of 60 statements. Responses to the statements are scored on a five-point Likert scale**

**1 (strongly agree), 2 (agree), 3 (neutral), 4 (disagree) and 5 (strongly disagree)**

1- I am not a worrier

2- I like to have a lot of people around me

3- I enjoy concentrating on a fantasy or daydream and exploring all its possibilities, letting it grow and develop

4- I try to be courteous to everyone I meet

5- I keep my belongings neat and clean

6- At times I have felt bitter and resentful

7- I laugh easily

8- I think it's interesting to learn and develop new hobbies

9- At times I bully or flatter people into doing what I want them to

10- I'm pretty good at pacing myself so as to get things done on time

11- When I'm under a great deal of stress sometimes I feel like I'm going to pieces

12- I prefer jobs that let me work alone without being bothered by other people

13- I'm intrigued by the patterns I find in art and nature

14- Some people think I'm selfish and egotistical

15- I often come into situations without being fully prepared

16- I rarely feel lonely or blue

17- I really enjoy talking to people

18- I believe letting students hear controversial speakers can only confuse and mislead them

19- If someone starts a fight, I'm ready to fight back

20- I try to perform all the tasks assigned to me conscientiously

21- I often feel tense and jittery

22- I like to be where the action is

23- Poetry has little or no effect on me

24- I'm better than most people and I know it

25- I have a clear set of goals and work toward them in an orderly fashion

26- Sometimes I feel completely worthless

27- I shy away from crowds of people

28- I would have difficulty just letting my mind wander without control or guidance

29- When I've been insulted, I just try to forgive and forget

30- I waste a lot of time before settling down to work

31- I rarely feel fearful or anxious

32- I often feel as if I'm bursting with energy

33- I seldom notice the moods or feelings that different environments produce

34- I tend to assume the best about people

35- I work hard to accomplish my goals

36- I often get angry at the way people treat me

37- I am a cheerful high-spirited person

38- I experience a wide range of emotions or feelings

39- Some people think of me as cold and calculating

40- When I make a commitment, I can always be counted on to follow through

41- Too often when things go wrong, I get discouraged and feel like giving up

42- I don’t get much pleasure from chatting with people

43- Sometimes when I am reading poetry or looking at awork of art I feel a chill or wave of excitement

44- I have no sympathy for beggars

45- Sometimes I’m not as dependable or reliable as I should be

46- I am seldom sad or depressed

47- My life is fast paced

48- I have little interest in speculating on the nature of the universe or the human condition

49- I generally try to be thoughtful and considerate

50- I am a productive person who always gets the job done

51- I often feel helpless and want someone else to solve my problems

52- I am a very active person

53- I have a lot of intellectual curiosity

54- If I don’t like people, I let them know it

55- I never seem to able to get organized

56- At times I have been so ashamed I just want to hide

57- I would rather go on my own way than be a leader of others

58- I often enjoy playing with theories or abstract ideas

59- If necessary, I am willing to manipulate people to get what I want

60- I strive for excellence in everything I do

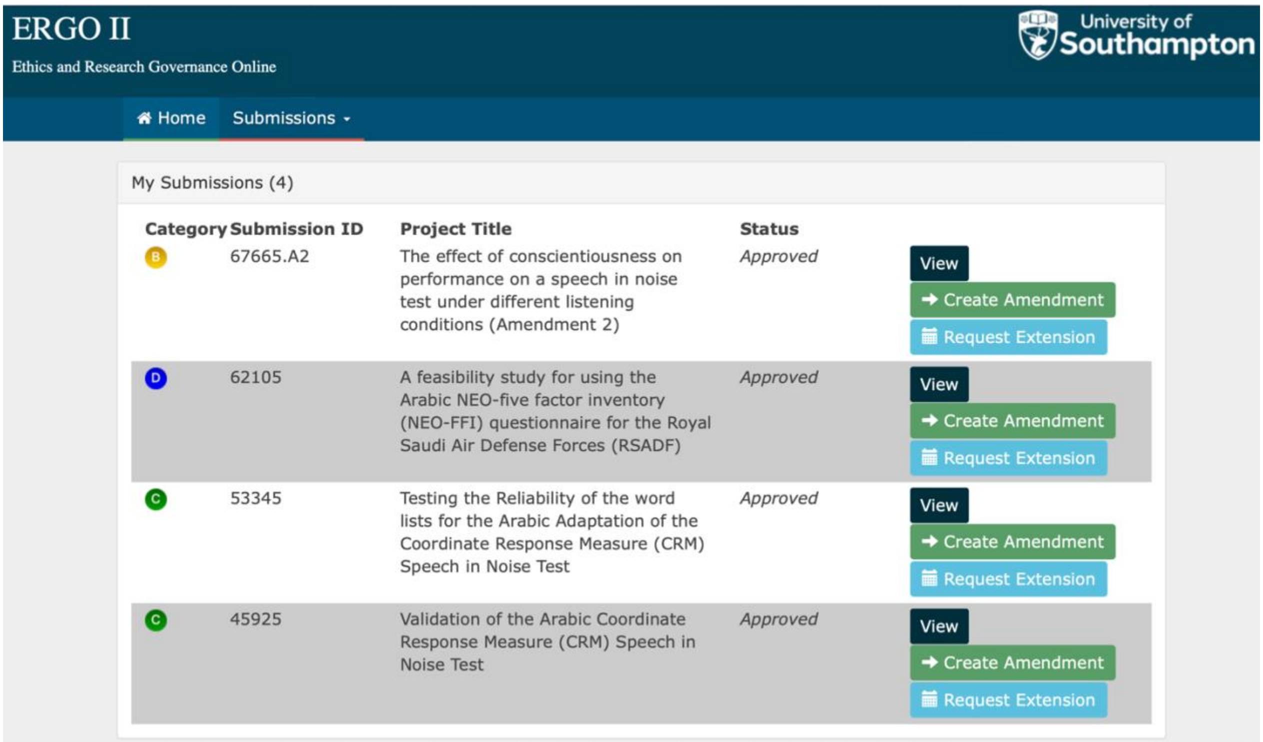
**Source:**

McCrae, R. and Costa, P., 2010. *NEO Inventories For The NEO Personality Inventory-3 (NEO-PI-3) NEO Five-Factor Inventory-3 (NEO-FFI-3) NEO Personality Inventory-Reviewed (NEO PI-R)*. Lutz, FL: PAR.

Attached below is a link to the Arabic version of the combined PIS & consent form attached to the study questionnaire. The consent from must be read and consent given in order for the study questionnaire to appear

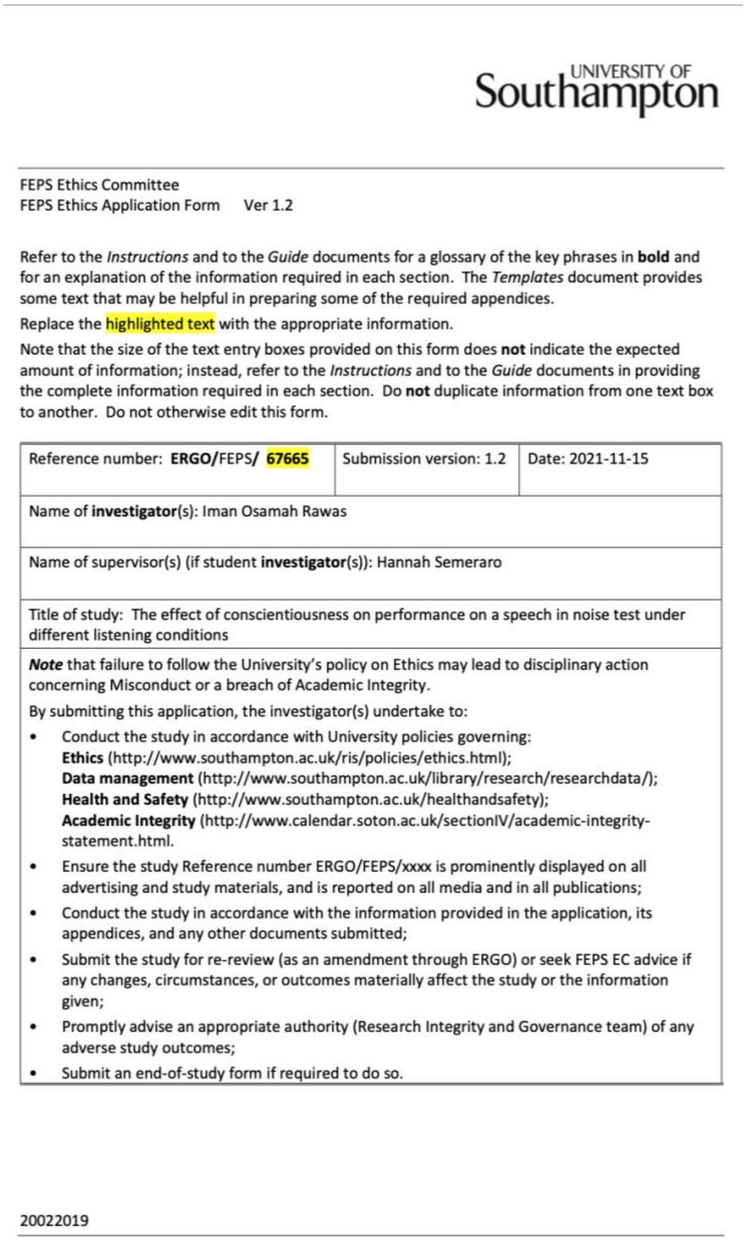
https://docs.google.com/forms/d/e/1FAIpQLSeXl8ir4Fuyx99-8sN_4xGa8M77gcf19-mbSnb2IOYxEkYc4A/viewform?usp=sf_link

# Appendix C University of Southampton’s Ethics and Research Governance Approvals


ERGO II
Ethics and Research Governance Online
University of Southampton
Home
Submissions
View
Create Amendment
Request Extension

My Submissions (4)

| Category | Submission ID | Project Title | Status |
|---|---|---|---|
| B | 67665.A2 | The effect of conscientiousness on performance on a speech in noise test under different listening conditions (Amendment 2) | *Approved* |
| D | 62105 | A feasibility study for using the Arabic NEO-five factor inventory (NEO-FFI) questionnaire for the Royal Saudi Air Defense Forces (RSADF) | *Approved* |
| C | 53345 | Testing the Reliability of the word lists for the Arabic Adaptation of the Coordinate Response Measure (CRM) Speech in Noise Test | *Approved* |
| C | 45925 | Validation of the Arabic Coordinate Response Measure (CRM) Speech in Noise Test | *Approved* |



UNIVERSITY OF Southampton

FEPS Ethics Committee
FEPS Ethics Application Form Ver 1.2

Refer to the *Instructions* and to the *Guide* documents for a glossary of the key phrases in **bold** and for an explanation of the information required in each section. The *Templates* document provides some text that may be helpful in preparing some of the required appendices.

Replace the highlighted text with the appropriate information.

Note that the size of the text entry boxes provided on this form does **not** indicate the expected amount of information; instead, refer to the *Instructions* and to the *Guide* documents in providing the complete information required in each section. Do **not** duplicate information from one text box to another. Do not otherwise edit this form.

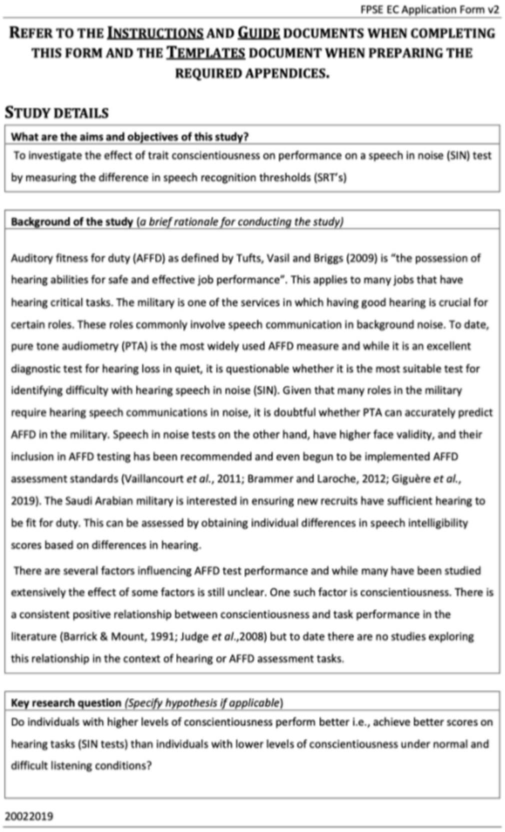
| | | |
|---|---|---|
| Reference number: **ERGO/**FEPS/ **67665** | Submission version: 1.2 | Date: 2021-11-15 |
| Name of **investigator**(s): Iman Osamah Rawas | | |
| Name of supervisor(s) (if student **investigator**(s)): Hannah Semeraro | | |
| Title of study: The effect of conscientiousness on performance on a speech in noise test under different listening conditions | | |

***Note*** that failure to follow the University's policy on Ethics may lead to disciplinary action concerning Misconduct or a breach of Academic Integrity.

By submitting this application, the investigator(s) undertake to:

- Conduct the study in accordance with University policies governing:
  **Ethics** (http://www.southampton.ac.uk/ris/policies/ethics.html);
  **Data management** (http://www.southampton.ac.uk/library/research/researchdata/);
  **Health and Safety** (http://www.southampton.ac.uk/healthandsafety);
  **Academic Integrity** (http://www.calendar.soton.ac.uk/sectionIV/academic-integrity-statement.html.
- Ensure the study Reference number ERGO/FEPS/xxxx is prominently displayed on all advertising and study materials, and is reported on all media and in all publications;
- Conduct the study in accordance with the information provided in the application, its appendices, and any other documents submitted;
- Submit the study for re-review (as an amendment through ERGO) or seek FEPS EC advice if any changes, circumstances, or outcomes materially affect the study or the information given;
- Promptly advise an appropriate authority (Research Integrity and Governance team) of any adverse study outcomes;
- Submit an end-of-study form if required to do so.

20022019



**REFER TO THE <u>INSTRUCTIONS</u> AND <u>GUIDE</u> DOCUMENTS WHEN COMPLETING THIS FORM AND THE <u>TEMPLATES</u> DOCUMENT WHEN PREPARING THE REQUIRED APPENDICES.**

## STUDY DETAILS

**What are the aims and objectives of this study?**

To investigate the effect of trait conscientiousness on performance on a speech in noise (SIN) test by measuring the difference in speech recognition thresholds (SRT's)

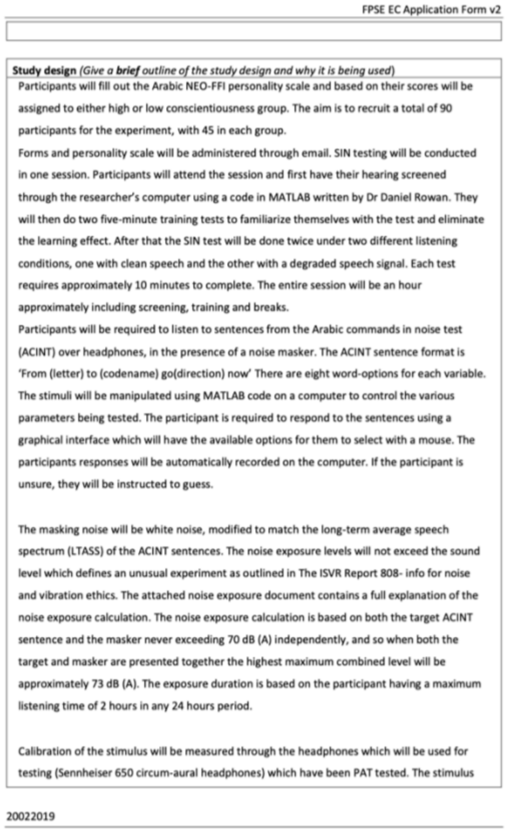
**Background of the study** *(a brief rationale for conducting the study)*

Auditory fitness for duty (AFFD) as defined by Tufts, Vasil and Briggs (2009) is "the possession of hearing abilities for safe and effective job performance". This applies to many jobs that have hearing critical tasks. The military is one of the services in which having good hearing is crucial for certain roles. These roles commonly involve speech communication in background noise. To date, pure tone audiometry (PTA) is the most widely used AFFD measure and while it is an excellent diagnostic test for hearing loss in quiet, it is questionable whether it is the most suitable test for identifying difficulty with hearing speech in noise (SIN). Given that many roles in the military require hearing speech communications in noise, it is doubtful whether PTA can accurately predict AFFD in the military. Speech in noise tests on the other hand, have higher face validity, and their inclusion in AFFD testing has been recommended and even begun to be implemented AFFD assessment standards (Vaillancourt *et al.*, 2011; Brammer and Laroche, 2012; Giguère *et al.*, 2019). The Saudi Arabian military is interested in ensuring new recruits have sufficient hearing to be fit for duty. This can be assessed by obtaining individual differences in speech intelligibility scores based on differences in hearing.

There are several factors influencing AFFD test performance and while many have been studied extensively the effect of some factors is still unclear. One such factor is conscientiousness. There is a consistent positive relationship between conscientiousness and task performance in the literature (Barrick & Mount, 1991; Judge *et al.*,2008) but to date there are no studies exploring this relationship in the context of hearing or AFFD assessment tasks.

**Key research question** *(Specify hypothesis if applicable*)

Do individuals with higher levels of conscientiousness perform better i.e., achieve better scores on hearing tasks (SIN tests) than individuals with lower levels of conscientiousness under normal and difficult listening conditions?



**Study design** *(Give a **brief** outline of the study design and why it is being used*)

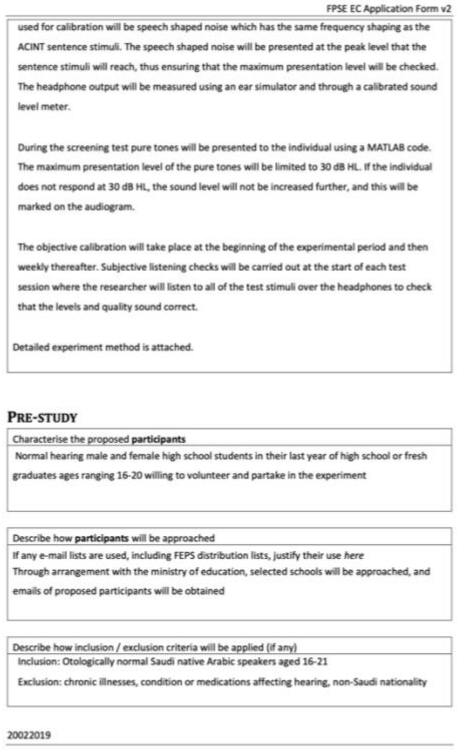
Participants will fill out the Arabic NEO-FFI personality scale and based on their scores will be assigned to either high or low conscientiousness group. The aim is to recruit a total of 90 participants for the experiment, with 45 in each group.

Forms and personality scale will be administered through email. SIN testing will be conducted in one session. Participants will attend the session and first have their hearing screened through the researcher's computer using a code in MATLAB written by Dr Daniel Rowan. They will then do two five-minute training tests to familiarize themselves with the test and eliminate the learning effect. After that the SIN test will be done twice under two different listening conditions, one with clean speech and the other with a degraded speech signal. Each test requires approximately 10 minutes to complete. The entire session will be an hour approximately including screening, training and breaks.

Participants will be required to listen to sentences from the Arabic commands in noise test (ACINT) over headphones, in the presence of a noise masker. The ACINT sentence format is 'From (letter) to (codename) go(direction) now' There are eight word-options for each variable. The stimuli will be manipulated using MATLAB code on a computer to control the various parameters being tested. The participant is required to respond to the sentences using a graphical interface which will have the available options for them to select with a mouse. The participants responses will be automatically recorded on the computer. If the participant is unsure, they will be instructed to guess.

The masking noise will be white noise, modified to match the long-term average speech spectrum (LTASS) of the ACINT sentences. The noise exposure levels will not exceed the sound level which defines an unusual experiment as outlined in The ISVR Report 808- info for noise and vibration ethics. The attached noise exposure document contains a full explanation of the noise exposure calculation. The noise exposure calculation is based on both the target ACINT sentence and the masker never exceeding 70 dB (A) independently, and so when both the target and masker are presented together the highest maximum combined level will be approximately 73 dB (A). The exposure duration is based on the participant having a maximum listening time of 2 hours in any 24 hours period.

Calibration of the stimulus will be measured through the headphones which will be used for testing (Sennheiser 650 circum-aural headphones) which have been PAT tested. The stimulus



used for calibration will be speech shaped noise which has the same frequency shaping as the ACINT sentence stimuli. The speech shaped noise will be presented at the peak level that the sentence stimuli will reach, thus ensuring that the maximum presentation level will be checked. The headphone output will be measured using an ear simulator and through a calibrated sound level meter.

During the screening test pure tones will be presented to the individual using a MATLAB code. The maximum presentation level of the pure tones will be limited to 30 dB HL. If the individual does not respond at 30 dB HL, the sound level will not be increased further, and this will be marked on the audiogram.

The objective calibration will take place at the beginning of the experimental period and then weekly thereafter. Subjective listening checks will be carried out at the start of each test session where the researcher will listen to all of the test stimuli over the headphones to check that the levels and quality sound correct.

Detailed experiment method is attached.

## PRE-STUDY

Characterise the proposed **participants**

Normal hearing male and female high school students in their last year of high school or fresh graduates ages ranging 16-20 willing to volunteer and partake in the experiment

Describe how **participants** will be approached

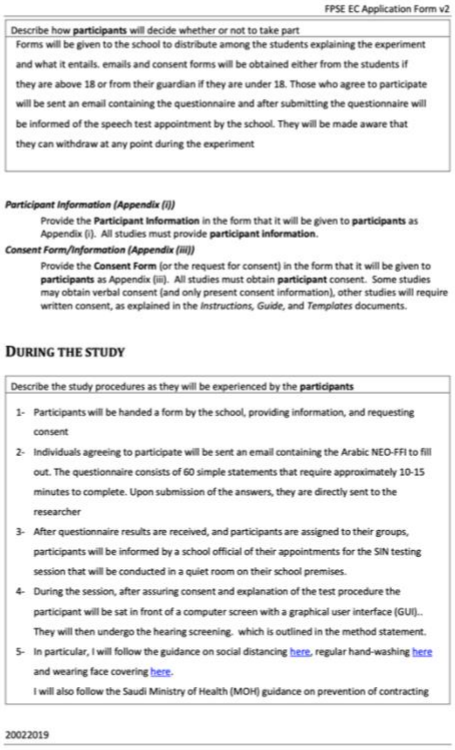
If any e-mail lists are used, including FEPS distribution lists, justify their use *here*

Through arrangement with the ministry of education, selected schools will be approached, and emails of proposed participants will be obtained

Describe how inclusion / exclusion criteria will be applied (if any)

Inclusion: Otologically normal Saudi native Arabic speakers aged 16-21

Exclusion: chronic illnesses, condition or medications affecting hearing, non-Saudi nationality



Describe how **participants** will decide whether or not to take part

Forms will be given to the school to distribute among the students explaining the experiment and what it entails. emails and consent forms will be obtained either from the students if they are above 18 or from their guardian if they are under 18. Those who agree to participate will be sent an email containing the questionnaire and after submitting the questionnaire will be informed of the speech test appointment by the school. They will be made aware that they can withdraw at any point during the experiment

***Participant Information (Appendix (i))***

Provide the **Participant Information** in the form that it will be given to **participants** as Appendix (i). All studies must provide **participant information**.

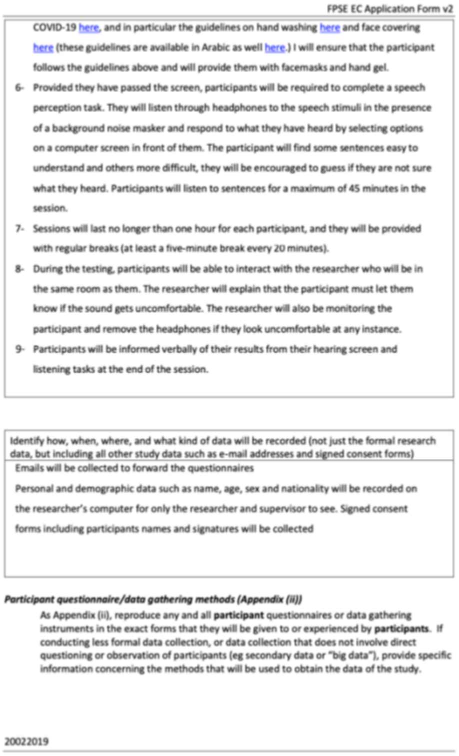
***Consent Form/Information (Appendix (iii))***

Provide the **Consent Form** (or the request for consent) in the form that it will be given to **participants** as Appendix (iii). All studies must obtain **participant** consent. Some studies may obtain verbal consent (and only present consent information), other studies will require written consent, as explained in the *Instructions, Guide*, and *Templates* documents.

## DURING THE STUDY

Describe the study procedures as they will be experienced by the **participants**

1- Participants will be handed a form by the school, providing information, and requesting consent
2- Individuals agreeing to participate will be sent an email containing the Arabic NEO-FFI to fill out. The questionnaire consists of 60 simple statements that require approximately 10-15 minutes to complete. Upon submission of the answers, they are directly sent to the researcher
3- After questionnaire results are received, and participants are assigned to their groups, participants will be informed by a school official of their appointments for the SIN testing session that will be conducted in a quiet room on their school premises.
4- During the session, after assuring consent and explanation of the test procedure the participant will be sat in front of a computer screen with a graphical user interface (GUI).. They will then undergo the hearing screening. which is outlined in the method statement.
5- In particular, I will follow the guidance on social distancing here, regular hand-washing here and wearing face covering here.
   I will also follow the Saudi Ministry of Health (MOH) guidance on prevention of contracting



COVID-19 here, and in particular the guidelines on hand washing here and face covering here (these guidelines are available in Arabic as well here.) I will ensure that the participant follows the guidelines above and will provide them with facemasks and hand gel.

6- Provided they have passed the screen, participants will be required to complete a speech perception task. They will listen through headphones to the speech stimuli in the presence of a background noise masker and respond to what they have heard by selecting options on a computer screen in front of them. The participant will find some sentences easy to understand and others more difficult, they will be encouraged to guess if they are not sure what they heard. Participants will listen to sentences for a maximum of 45 minutes in the session.

7- Sessions will last no longer than one hour for each participant, and they will be provided with regular breaks (at least a five-minute break every 20 minutes).

8- During the testing, participants will be able to interact with the researcher who will be in the same room as them. The researcher will explain that the participant must let them know if the sound gets uncomfortable. The researcher will also be monitoring the participant and remove the headphones if they look uncomfortable at any instance.

9- Participants will be informed verbally of their results from their hearing screen and listening tasks at the end of the session.

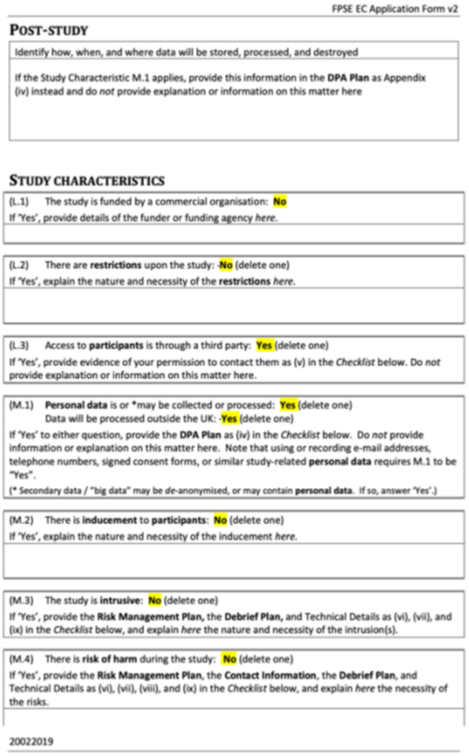
Identify how, when, where, and what kind of data will be recorded (not just the formal research data, but including all other study data such as e-mail addresses and signed consent forms)

Emails will be collected to forward the questionnaires

Personal and demographic data such as name, age, sex and nationality will be recorded on the researcher's computer for only the researcher and supervisor to see. Signed consent forms including participants names and signatures will be collected

***Participant questionnaire/data gathering methods (Appendix (ii))***

As Appendix (ii), reproduce any and all **participant** questionnaires or data gathering instruments in the exact forms that they will be given to or experienced by **participants**. If conducting less formal data collection, or data collection that does not involve direct questioning or observation of participants (eg secondary data or "big data"), provide specific information concerning the methods that will be used to obtain the data of the study.



## POST-STUDY

Identify how, when, and where data will be stored, processed, and destroyed

If the Study Characteristic M.1 applies, provide this information in the **DPA Plan** as Appendix (iv) instead and do *not* provide explanation or information on this matter here

## STUDY CHARACTERISTICS

(L.1) The study is funded by a commercial organisation: **No**
If 'Yes', provide details of the funder or funding agency *here*.

(L.2) There are **restrictions** upon the study: -**No** (delete one)
If 'Yes', explain the nature and necessity of the **restrictions** *here*.

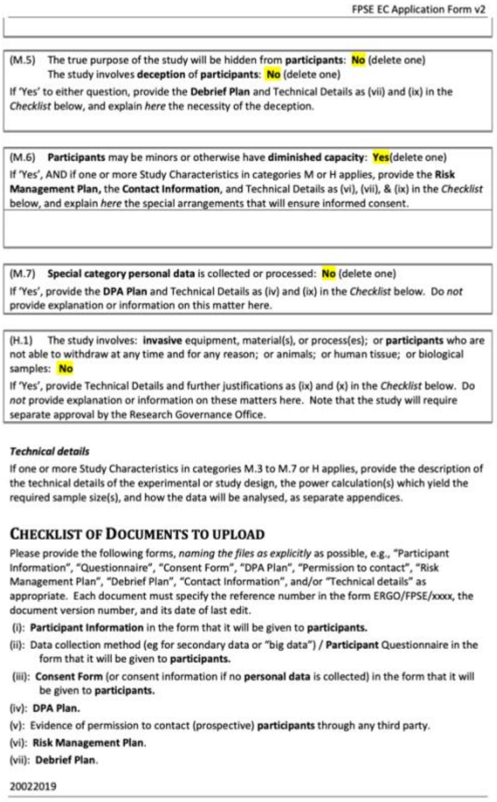
(L.3) Access to **participants** is through a third party: **Yes** (delete one)
If 'Yes', provide evidence of your permission to contact them as (v) in the *Checklist* below. Do *not* provide explanation or information on this matter here.

(M.1) **Personal data** is or *may be collected or processed: **Yes** (delete one)
Data will be processed outside the UK: -**Yes** (delete one)
If 'Yes' to either question, provide the **DPA Plan** as (iv) in the *Checklist* below. Do *not* provide information or explanation on this matter here. Note that using or recording e-mail addresses, telephone numbers, signed consent forms, or similar study-related **personal data** requires M.1 to be "Yes".
(* Secondary data / "big data" may be *de*-anonymised, or may contain **personal data**. If so, answer 'Yes'.)

(M.2) There is **inducement** to **participants**: **No** (delete one)
If 'Yes', explain the nature and necessity of the inducement *here*.

(M.3) The study is **intrusive**: **No** (delete one)
If 'Yes', provide the **Risk Management Plan,** the **Debrief Plan,** and Technical Details as (vi), (vii), and (ix) in the *Checklist* below, and explain *here* the nature and necessity of the intrusion(s).

(M.4) There is **risk of harm** during the study: **No** (delete one)
If 'Yes', provide the **Risk Management Plan**, the **Contact Information**, the **Debrief Plan**, and Technical Details as (vi), (vii), (viii), and (ix) in the *Checklist* below, and explain *here* the necessity of the risks.



(M.5) The true purpose of the study will be hidden from **participants**: **No** (delete one)
The study involves **deception** of **participants**: **No** (delete one)
If 'Yes' to either question, provide the **Debrief Plan** and Technical Details as (vii) and (ix) in the *Checklist* below, and explain *here* the necessity of the deception.

(M.6) **Participants** may be minors or otherwise have **diminished capacity**: **Yes**(delete one)
If 'Yes', AND if one or more Study Characteristics in categories M or H applies, provide the **Risk Management Plan,** the **Contact Information**, and Technical Details as (vi), (vii), & (ix) in the *Checklist* below, and explain *here* the special arrangements that will ensure informed consent.

(M.7) **Special category personal data** is collected or processed: **No** (delete one)
If 'Yes', provide the **DPA Plan** and Technical Details as (iv) and (ix) in the *Checklist* below. Do *not* provide explanation or information on this matter here.

(H.1) The study involves: **invasive** equipment, material(s), or process(es); or **participants** who are not able to withdraw at any time and for any reason; or animals; or human tissue; or biological samples: **No**
If 'Yes', provide Technical Details and further justifications as (ix) and (x) in the *Checklist* below. Do *not* provide explanation or information on these matters here. Note that the study will require separate approval by the Research Governance Office.

***Technical details***

If one or more Study Characteristics in categories M.3 to M.7 or H applies, provide the description of the technical details of the experimental or study design, the power calculation(s) which yield the required sample size(s), and how the data will be analysed, as separate appendices.

## CHECKLIST OF DOCUMENTS TO UPLOAD

Please provide the following forms, *naming the files as explicitly* as possible, e.g., "Participant Information", "Questionnaire", "Consent Form", "DPA Plan", "Permission to contact", "Risk Management Plan", "Debrief Plan", "Contact Information", and/or "Technical details" as appropriate. Each document must specify the reference number in the form ERGO/FPSE/xxxx, the document version number, and its date of last edit.

(i): **Participant Information** in the form that it will be given to **participants.**

(ii): Data collection method (eg for secondary data or "big data") / **Participant** Questionnaire in the form that it will be given to **participants.**

(iii): **Consent Form** (or consent information if no **personal data** is collected) in the form that it will be given to **participants.**

(iv): **DPA Plan.**

(v): Evidence of permission to contact (prospective) **participants** through any third party.

(vi): **Risk Management Plan.**

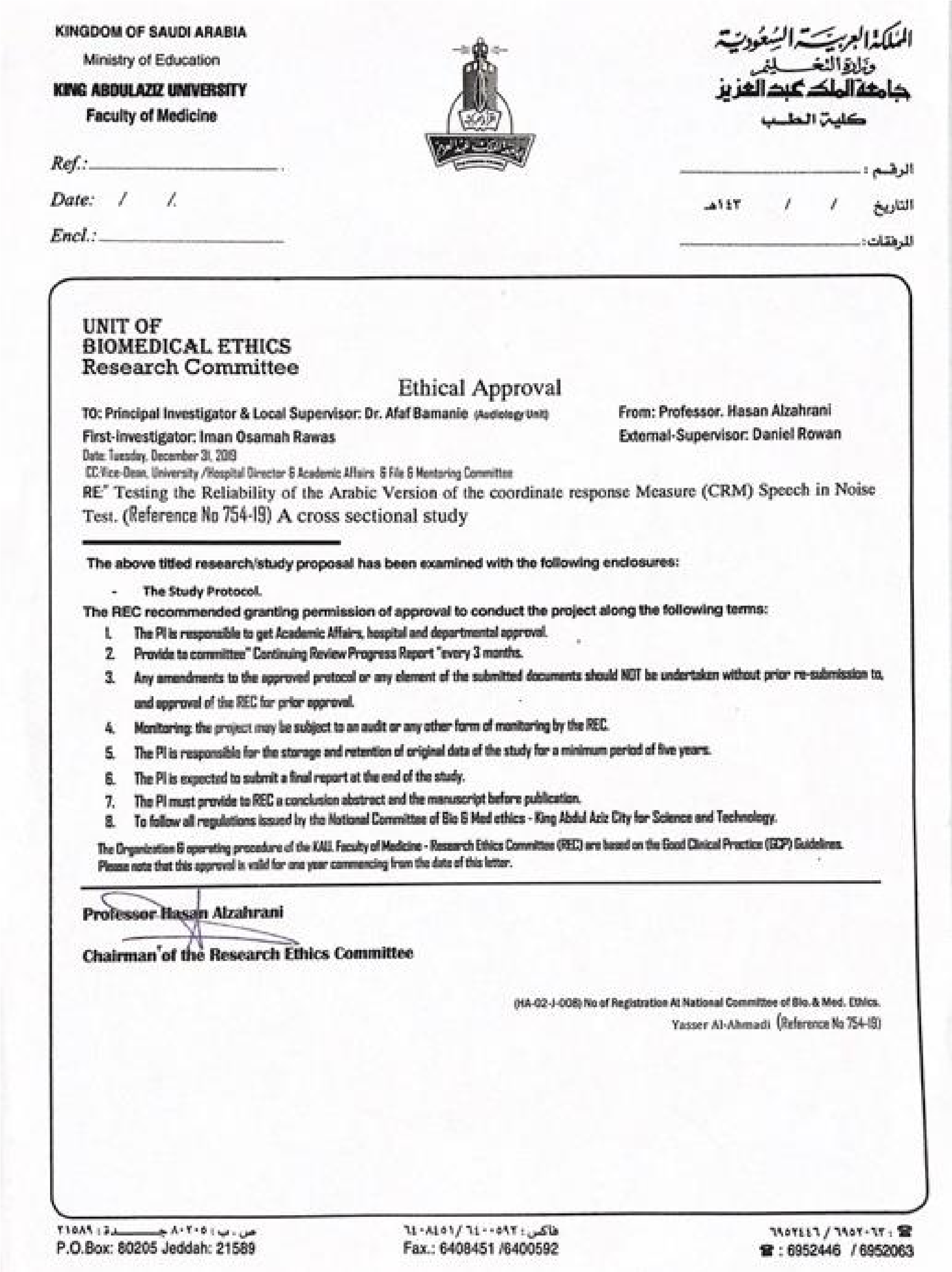
(vii): **Debrief Plan.**

(viii): **Contact Information**.

(ix): Technical details of the experimental or study design, the power calculation(s) for the required sample size(s), and how the data will be analysed.

(x): Further details and justifications in the case of: **invasive** equipment, material(s), or process(es); **participants** who are not able to withdraw at any time and for any reason; animals; human tissue; or biological samples.

# Appendix D King Abdul-Aziz University Research ethics committee’s approval for studies one, two and five



KINGDOM OF SAUDI ARABIA
Ministry of Education
KING ABDULAZIZ UNIVERSITY
Faculty of Medicine

المملكة العربية السعودية
وزارة التعليم
جامعة الملك عبدالعزيز
كلية الطب

Ref.:

Date: / /

Encl.:

الرقم :

التاريخ / / ١٤٣ هـ

المرفقات:

UNIT OF
BIOMEDICAL ETHICS
Research Committee

Ethical Approval

TO: Principal Investigator & Local Supervisor: Dr. Afaf Bamanie (Audiology Unit)
First-investigator: Iman Osamah Rawas

From: Professor. Hasan Alzahrani
External-Supervisor: Daniel Rowan

Date: Tuesday, December 31, 2019
CC:Vice-Dean, University /Hospital Director & Academic Affairs & File & Mentoring Committee

RE" Testing the Reliability of the Arabic Version of the coordinate response Measure (CRM) Speech in Noise Test. (Reference No 754-19) A cross sectional study

The above titled research/study proposal has been examined with the following enclosures:

- The Study Protocol.

The REC recommended granting permission of approval to conduct the project along the following terms:

1. The PI is responsible to get Academic Affairs, hospital and departmental approval.
2. Provide to committee" Continuing Review Progress Report "every 3 months.
3. Any amendments to the approved protocol or any element of the submitted documents should NOT be undertaken without prior re-submission to, and approval of the REC for prior approval.
4. Monitoring: the project may be subject to an audit or any other form of monitoring by the REC.
5. The PI is responsible for the storage and retention of original data of the study for a minimum period of five years.
6. The PI is expected to submit a final report at the end of the study.
7. The PI must provide to REC a conclusion abstract and the manuscript before publication.
8. To follow all regulations issued by the National Committee of Bio & Med ethics - King Abdul Aziz City for Science and Technology.

The Organization & operating procedure of the KAU. Faculty of Medicine - Research Ethics Committee (REC) are based on the Good Clinical Practice (GCP) Guidelines.
Please note that this approval is valid for one year commencing from the date of this letter.

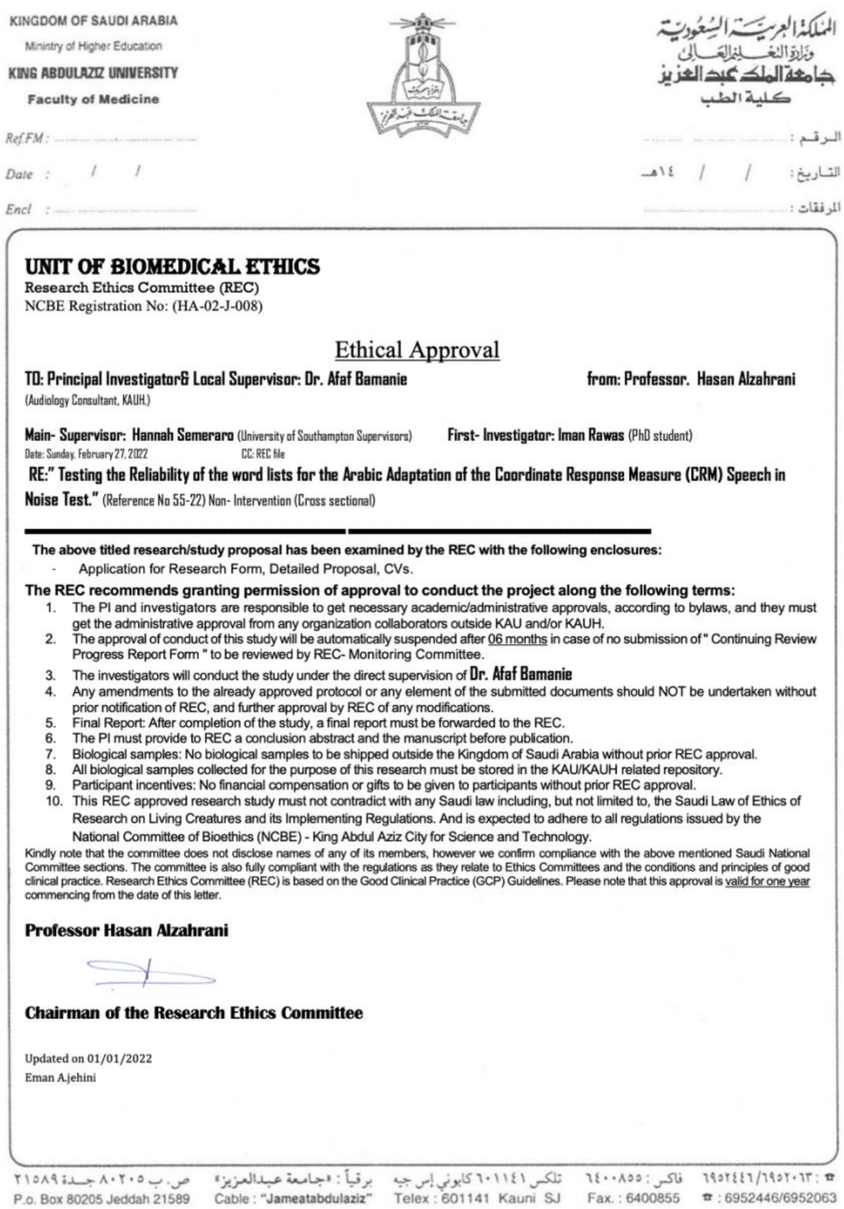
Professor Hasan Alzahrani
Chairman of the Research Ethics Committee

(HA-02-J-008) No of Registration At National Committee of Bio.& Med. Ethics.
Yasser Al-Ahmadi (Reference No 754-19)

ص . ب : ٨٠٢٠٥ جـدة : ٢١٥٨٩
P.O.Box: 80205 Jeddah: 21589

فاكس: ٦٤٠٠٥٩٢ / ٦٤٠٨٤٥١
Fax.: 6408451 /6400592

☎ : ٦٩٥٢٠٦٣ / ٦٩٥٢٤٤٦
☎ : 6952446 / 6952063



KINGDOM OF SAUDI ARABIA
Minintry of Higher Education
KING ABDULAZIZ UNIVERSITY
Faculty of Medicine

المملكة العربية السعودية
وزارة التعليم العالي
جامعة الملك عبد العزيز
كلية الطب

Ref.FM : الرقم :

Date : / / التاريخ : / / ١٤هـ

Encl : المرفقات :

**UNIT OF BIOMEDICAL ETHICS**
**Research Ethics Committee (REC)**
NCBE Registration No: (HA-02-J-008)

Ethical Approval

**TO: Principal Investigator& Local Supervisor: Dr. Afaf Bamanie** **from: Professor. Hasan Alzahrani**
(Audiology Consultant, KAUH.)

**Main- Supervisor: Hannah Semeraro** (University of Southampton Supervisors) **First- Investigator: Iman Rawas** (PhD student)
Date: Sunday, February 27, 2022 CC: REC file

**RE:" Testing the Reliability of the word lists for the Arabic Adaptation of the Coordinate Response Measure (CRM) Speech in Noise Test."** (Reference No 55-22) Non- Intervention (Cross sectional)

**The above titled research/study proposal has been examined by the REC with the following enclosures:**

- Application for Research Form, Detailed Proposal, CVs.

**The REC recommends granting permission of approval to conduct the project along the following terms:**

1. The PI and investigators are responsible to get necessary academic/administrative approvals, according to bylaws, and they must get the administrative approval from any organization collaborators outside KAU and/or KAUH.
2. The approval of conduct of this study will be automatically suspended after 06 months in case of no submission of " Continuing Review Progress Report Form " to be reviewed by REC- Monitoring Committee.
3. The investigators will conduct the study under the direct supervision of **Dr. Afaf Bamanie**
4. Any amendments to the already approved protocol or any element of the submitted documents should NOT be undertaken without prior notification of REC, and further approval by REC of any modifications.
5. Final Report: After completion of the study, a final report must be forwarded to the REC.
6. The PI must provide to REC a conclusion abstract and the manuscript before publication.
7. Biological samples: No biological samples to be shipped outside the Kingdom of Saudi Arabia without prior REC approval.
8. All biological samples collected for the purpose of this research must be stored in the KAU/KAUH related repository.
9. Participant incentives: No financial compensation or gifts to be given to participants without prior REC approval.
10. This REC approved research study must not contradict with any Saudi law including, but not limited to, the Saudi Law of Ethics of Research on Living Creatures and its Implementing Regulations. And is expected to adhere to all regulations issued by the National Committee of Bioethics (NCBE) - King Abdul Aziz City for Science and Technology.

Kindly note that the committee does not disclose names of any of its members, however we confirm compliance with the above mentioned Saudi National Committee sections. The committee is also fully compliant with the regulations as they relate to Ethics Committees and the conditions and principles of good clinical practice. Research Ethics Committee (REC) is based on the Good Clinical Practice (GCP) Guidelines. Please note that this approval is valid for one year commencing from the date of this letter.

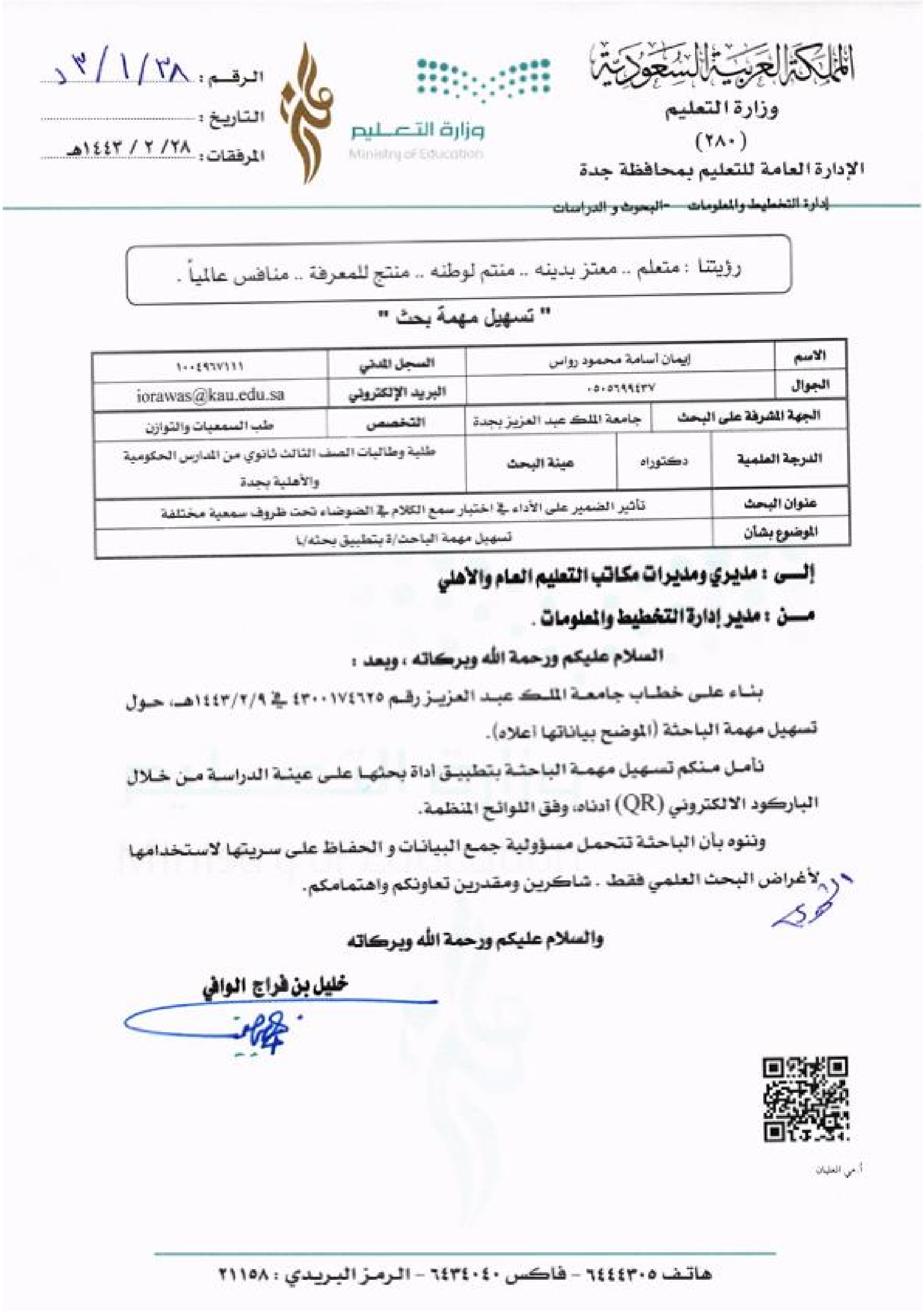
**Professor Hasan Alzahrani**

**Chairman of the Research Ethics Committee**

Updated on 01/01/2022
Eman A.jehini

ص. ب ٨٠٢٠٥ جـدة ٢١٥٨٩ برقياً : «جامعة عبدالعزيز» تلكس ٦٠١١٤١ كايوني إس جيه فاكس : ٦٤٠٠٨٥٥ ☎ : ٦٩٥٢٤٤٦/٦٩٥٢٠٦٣
P.o. Box 80205 Jeddah 21589 Cable : "Jameatabdulaziz" Telex : 601141 Kauni SJ Fax. : 6400855 ☎ : 6952446/6952063

# Appendix E Ministry of Education’s approval for study five



المملكة العربية السعودية
وزارة التعليم
(٢٨٠)
الإدارة العامة للتعليم بمحافظة جدة
إدارة التخطيط والمعلومات –البحوث و الدراسات

وزارة التعليم
Ministry of Education

الرقم : ٣٨/١/٣د
التاريخ :
المرفقات : ٢٨/ ٢ /١٤٤٣هـ

رؤيتنا : متعلم .. معتز بدينه .. منتم لوطنه .. منتج للمعرفة .. منافس عالمياً .

" تسهيل مهمة بحث "

| | | | |
|---|---|---|---|
| الاسم | إيمان أسامة محمود رواس | السجل المدني | ١٠٠٤٩٦٧١١١ |
| الجوال | ٠٥٠٥٦٩٩٤٣٧ | البريد الإلكتروني | iorawas@kau.edu.sa |
| الجهة المشرفة على البحث | جامعة الملك عبد العزيز بجدة | التخصص | طب السمعيات والتوازن |
| الدرجة العلمية | دكتوراه | عينة البحث | طلبة وطالبات الصف الثالث ثانوي من المدارس الحكومية والأهلية بجدة |
| عنوان البحث | تأثير الضمير على الأداء في اختبار سمع الكلام في الضوضاء تحت ظروف سمعية مختلفة | | |
| الموضوع بشأن | تسهيل مهمة الباحث/ة بتطبيق بحثه/ا | | |

إلـــى : مديري ومديرات مكاتب التعليم العام والأهلي

مـــن : مدير إدارة التخطيط والمعلومات .

السلام عليكم ورحمة الله وبركاته ، وبعد :

بناء على خطاب جامعة الملك عبد العزيز رقم ٤٣٠٠١٧٤٦٢٥ في ١٤٤٣/٢/٩هـ، حول تسهيل مهمة الباحثة (الموضح بياناتها اعلاه).

نأمل منكم تسهيل مهمة الباحثة بتطبيق اداة بحثها على عينة الدراسة من خلال الباركود الالكتروني (QR) أدناه، وفق اللوائح المنظمة.

وننوه بأن الباحثة تتحمل مسؤولية جمع البيانات و الحفاظ على سريتها لاستخدامها لأغراض البحث العلمي فقط . شاكرين ومقدرين تعاونكم واهتمامكم.

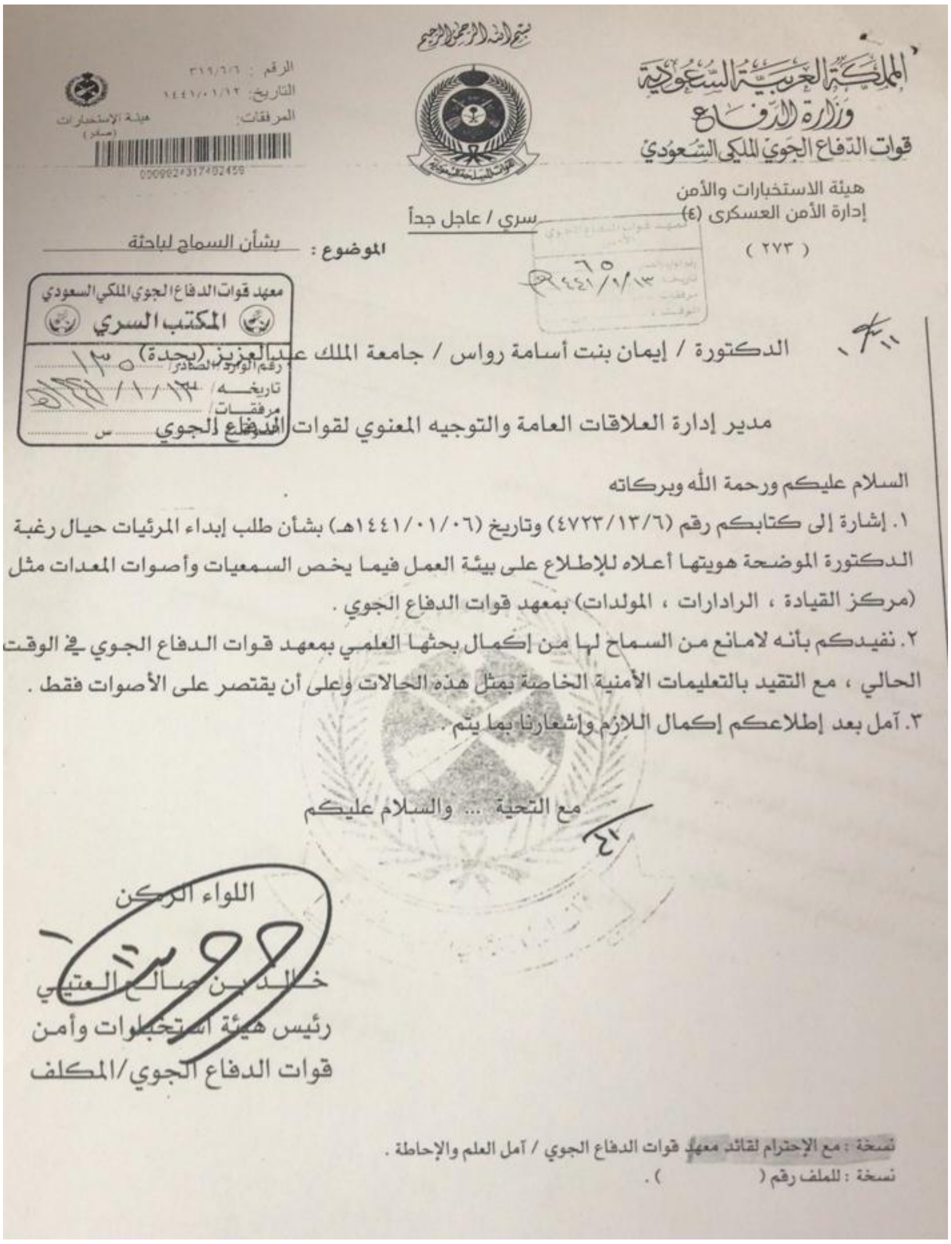
والسلام عليكم ورحمة الله وبركاته

خليل بن فراج الوافي

أ. مي العليان

هاتف ٦٤٤٤٣٠٥ – فاكس ٦٤٣٤٠٤٠ – الرمز البريدي : ٢١١٥٨

# Appendix F Royal Saudi Air Defence’s approval for study four



بسم الله الرحمن الرحيم

المملكة العربية السعودية
وزارة الدفاع
قوات الدفاع الجوي الملكي السعودي

هيئة الاستخبارات والأمن
إدارة الأمن العسكرى (٤)
( ٢٧٣ )

الرقم : ٣١٩/٦/٦
التاريخ: ١٤٤١/٠١/١٢
المرفقات:

هيئة الإستخبارات
(صادر)

000982#317402459

سري / عاجل جداً

الموضوع : بشأن السماح لباحثة

٦٥
١٤٤١/١/١٣

معهد قوات الدفاع الجوي الملكي السعودي
المكتب السري
رقم الوارد/الصادر ١٣٠٥
تاريخه/ ١/١/١٣
مرفقات/
س

الدكتورة / إيمان بنت أسامة رواس / جامعة الملك عبدالعزيز (بجدة)

مدير إدارة العلاقات العامة والتوجيه المعنوي لقوات الدفاع الجوي

السلام عليكم ورحمة الله وبركاته

١. إشارة إلى كتابكم رقم (٤٧٢٣/١٣/٦) وتاريخ (١٤٤١/٠١/٠٦هـ) بشأن طلب إبداء المرئيات حيال رغبة الدكتورة الموضحة هويتها أعلاه للإطلاع على بيئة العمل فيما يخص السمعيات وأصوات المعدات مثل (مركز القيادة ، الرادارات ، المولدات) بمعهد قوات الدفاع الجوي .

٢. نفيدكم بأنه لامانع من السماح لها من إكمال بحثها العلمي بمعهد قوات الدفاع الجوي في الوقت الحالي ، مع التقيد بالتعليمات الأمنية الخاصة بمثل هذه الحالات وعلى أن يقتصر على الأصوات فقط .

٣. آمل بعد إطلاعكم إكمال اللازم وإشعارنا بما يتم .

مع التحية ... والسلام عليكم

اللواء الركن
خالد بن صالح العتيبي
رئيس هيئة استخبارات وأمن
قوات الدفاع الجوي/المكلف

نسخة : مع الإحترام لقائد معهد قوات الدفاع الجوي / آمل العلم والإحاطة .
نسخة : للملف رقم ( ) .

# Appendix G Phonetic Analyses of speech test corpus and common military phrases

| Important/ Common military words and phrases | | | | | |
|---|---|---|---|---|---|
| التحليل الصوتي | الكلمة | التحليل الصوتي | الكلمة | التحليل الصوتي | الكلمة |
| /iʃtabɑk/ | اشتباك | /ðˤajr masmu:ʕ/ | غير مسموع | / ʒajɪd/ | جيد |
| /ɣ aτ̲τ̲ɩ/ | غطي | /ʔila:/ | إلى | /mafhʊm/ | مفهوم |
| /ʔiqtˤaʕ ʔɑdˤarb/ | قطع<br>الضرب | /ʔintɑðˤɪr/ | انتظر | /taʃwi:ʃ/ | تشويش |
| /ʔiqtˤaʕ ʔɑlɩʃtiba:k/ | قطع<br>الشباك | /ʔɑrsɩl/ | ارسل | /dˤʕi:f/ | ضعيف |
| /ʔɑbja: dˤ/ | أبيض | /ʔistɑlamt/ | استلمت | /maʕak/ | معك |
| /ʔɑsˤfar/ | أصفر | /tɑballaðˤτ/ | تبلغت | /jeτκɑlɑ:m/ | يكلم |
| /ʔaħmar/ | أحمر | /ʔɑʒɩb/ | أجب | /lɑwħa/ | لوحة |
| /ħo:r'/ | حر | /i:ntaha/ | انتهى | /ʔɑrqa:m/ | أرقام |
| /mufi:d/ | مفيد | /jeτκɑlɑ:m/ | يكلم | /ʔaħrʊf/ | أحرف |
| /mɑħuo:r/ | محظور | /sˤħiħ/ | صحيح | /naʕam/ | نعم |
| /masmu: ʕ/ | مسموع | /bɑrqɩja/ | برقية | /laʔ/ | لاء |
| | | /ʔaʕɪd/ | أعد | /tʌsˤ ħi: ħ/ | تصحيح |
| | | /ʕinwa:n/ | عنوان | /dɑrapαδZa/ | درجة |
| | | /nɩda:ʔ/ | نداء | /xaτ̲a̲ʔ/ | خطأ |
| | | /tɑħririjφa/ | تحريرية | /ʔɑsbaqɩjφa/ | أسبقية |

| Test words for the ACINT | | | | | |
|---|---|---|---|---|---|
| /ʔɩma:m/ | إمام | /fɑras/ | فرس | /ka:f/ | كاف |
| /batˤ:ʔ/ | بطيء | /nimɩr/ | نمر | /la:m/ | لام |
| / ʒɑnʊ:b/ | جنوب | /fɑhɑd/ | فهد | /mɩm/ | ميم |
| /sɑri:ʕ/ | سريع | /matˤar/ | مطر | /nʊn/ | نون |
| /qari:b/ | قريب | /ʒɑbal/ | جبل | /zi:n/ | زين |
| /baʕi:d/ | بعيد | /sˤaqr/ | صقر | /ʕa:n/ | عين |
| /ʕ a:li/ | عالي | /raʕd | رعد | /sˤa:d/ | صاد |
| /jami:n/ | يمين | /badr/ | بدر | /wɑ:w/ | واو |

# Appendix H Studies within the last ten years examining the conscientiousness- task performance relationship using the NEO inventory

| study | Outcome measured<br>Tools used | sample | takeaway |
|---|---|---|---|
| Differentiated measurement of conscientiousness and emotional stability in an occupational context–greater effort or greater benefit? (Moldzio,2021) | Using NEO-FFI, Their own test (ABGS) and measurements of cognitive ability<br>They measured C (industriousness and orderliness) and emotional stability (social interactive and continuous | 5972 | In a sub-sample of trainees with a commercial focus, a low positive relationship was found between the two aspects of conscientiousness and exam grades |
| The multiple face(t)s of state conscientiousness: Predicting task performance and organizational citizenship behavior (Debusscher et al, 2017) | C subscale of NEO-PI-R, a self-measure of task performance and a scale for measuring organizational citizenship behaviour | 83 | Global conscientiousness factor also significantly predicted within person daily task performance (b = 0.63, 95% CI [0.40, 0.87]) |
| Noncognitive Indicators as Critical Predictors of Students' Performance in Dental School (Stacey & Kurunathan, 2015) | NEO-PI-3 result correlated with grade point average (GPA), and patient management courses | 292 | All C facets were associated with higher GPAs and clinical scores ranging from (*r*= 0.29-0.39) |
| The five-factor model of personality traits and organizational citizenship behaviors: A meta-analysis (Chiaburu et al, 2011) | Meta-analysis of 77 studies | | Performance is associated more strongly with contextual aspects of job performance rather than task performance |

# List of References

الأنصاري،ب مدى كفاءة قائمة العوامل الخمسة الكبرى للشخصية في المجتمع الكويتي', *دراسات نفسية*, ٧(٢)، ٢٧٧-٣١٠ (١٩٩٧) ; Al-Ansari, B. (1997) 'The efficiency of the Five-Factor Inventory in a Kuwaiti population', *Psychological Studies,* 7(2), pp. 277-310.

Al-Omari, A. S. *et al.* (2018) ‘Association of flying time with hearing loss in military pilots’, *Saudi Journal of Medicine and Medical Sciences*, 6(3), p. 155. doi: 10.4103/sjmms.sjmms_10_18.

Al-Saif, S. S. (2014) ‘Screening for NIHL among military personnel in eastern province of Saudi Arabia’, *Otolaryngology–Head and Neck Surgery*, 151(1_suppl), pp. P226–P227. doi: 10.1177/0194599814541629a282.

Almutairi, S., Heller, M. and Yen, D. (2021) ‘Reclaiming the heterogeneity of the Arab states’, *Cross Cultural and Strategic Management*, 28(1), pp. 158–176. doi: 10.1108/CCSM-09-2019-0170.

Aluja, A. *et al.* (2005) ‘Comparison of the NEO-FFI, the NEO-FFI-R and an alternative short version of the NEO-PI-R (NEO-60) in Swiss and Spanish samples’, *Personality and Individual Differences*, 38(3), pp. 591–604. doi: 10.1016/j.paid.2004.05.014.

Alzoubi, A. M. and Alkmayseh, O. S. (2019) ‘The predictive power of big five personality traits and some variables in positivity among Albalqa Applied University students’, *Journal of Educational Sciences*, 1(2), pp. 339–362.

Anderson, S. *et al.* (2013) ‘A dynamic auditory-cognitive system supports speech-in-noise perception in older adults’, *Hearing Research*, 300, pp. 18–32. doi: 10.1016/j.heares.2013.03.006.

Anderson, S. and Kraus, N. (2010) ‘Sensory-cognitive interaction in the neural encoding of speech in noise: A review’, *Journal of the American Academy of Audiology*, 21(9), pp. 575–585. doi: 10.3766/jaaa.21.9.3.

Andreeva, I. G. (2018) ‘Spatial selectivity of hearing in speech recognition in speech-shaped noise environment’, *Human Physiology*, 44(2), pp. 226–236. doi: 10.1134/S0362119718020020.

Antonio, K. and Zhang, Y. (2014) *Modeling Applications in Actuarial Science*. Cambridge University Press. doi: 10.1017/CBO9781139342674.008.

Baddeley, A. (2003) ‘Working memory: looking back and looking forward.’ *Nature Reviews Neuroscience,* 4, pp. 829–839 (2003). https://doi.org/10.1038/nrn1201

Baer, T. and Moore, B. C. J. (1994) ‘Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech’, *The Journal of the Acoustical Society of America*, 95(4), pp. 2277–2280. doi: 10.1121/1.408640.

Baiduc, R. R. and Helzner, E. P. (2019) ‘Epidemiology of diabetes and hearing loss’, *Seminars in Hearing*, 40(4), pp. 281–291. doi: 10.1055/s-0039-1697643.

Bakker, A. B., Demerouti, E. and Ten Brummelhuis, L. L. (2012) ‘Work engagement, performance, and active learning: The role of conscientiousness’, *Journal of Vocational Behavior*. 80(2), pp. 555–564. doi: 10.1016/j.jvb.2011.08.008.

Barrenäs, M. L. and Wikström, I. (2000) ‘The influence of hearing and age on speech recognition scores in noise in audiological patients and in the general population’, *Ear and Hearing*, 21(6), pp. 569–577. doi: 10.1097/00003446-200012000-00004.

Bartlett, J. W. and Frost, C. (2008) ‘Reliability, repeatability and reproducibility: Analysis of measurement errors in continuous variables’, *Ultrasound in Obstetrics and Gynecology*, 31(4), pp. 466–475. doi: 10.1002/uog.5256.

Bauer, P., Jones, J. and Fingscheidt, T. (2013) ‘Impact of hearing impairment on fricative intelligibility for artificially bandwidth-extended telephone speech in noise’, *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. IEEE, pp. 7039–7043. doi: 10.1109/ICASSP.2013.6639027.

Bech, S. C., Dammeyer, J. and Liu, J. (2021) ‘Changes in personality traits among candidates for special operations forces’, *Military Psychology*. 33(3), pp. 197–204. doi: 10.1080/08995605.2021.1902178.

Bentler, R. A. (2000) ‘List equivalency and test-retest reliability of the Speech in Noise Test’, *American Journal of Audiology*, 9(2), pp. 84–100.

Bevis, Z. *et al.* (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. doi: 10.4103/1463-1741.132101.

Bickenbach, J. (2011) ‘The world report on disability’, *Disability and Society*. 26(5), pp. 655–658. doi: 10.1080/09687599.2011.589198.

Bidelman, G. M. and Momtaz, S. (2021) ‘Subcortical rather than cortical sources of the frequency-following response (FFR) relate to speech-in-noise perception in normal-hearing listeners’, *Neuroscience Letters*. 746(December 2020), p. 135664. doi: 10.1016/j.neulet.2021.135664.

Bilgiç, R. and Sümer, H. C. (2009) ‘Predicting military performance from specific personality measures: A validity study’, *International Journal of Selection and Assessment*, 17(2), pp. 231–238. doi: 10.1111/j.1468-2389.2009.00465.x.

Bipin Kishore, P. (2020) ‘Binaural hearing: Physiological and clinical view’, *Archives of Otolaryngology and Rhinology*, 6(2), pp. 033–036. doi: 10.17352/2455-1759.000118.

Bland JM, Altman DG. (1986) 'Statistical methods for assessing agreement between two methods of clinical measurement', *Lancet*, 1(8476), pp. 307-310.

Bobdey, S. *et al.* (2021) ‘Association of personality traits with performance in military training’, *Medical Journal Armed Forces India*. 77(4), pp. 431–436. doi: 10.1016/j.mjafi.2020.12.022.

Bolia, R. S. *et al.* (2000) ‘A speech corpus for multitalker communications research’, *The Journal of the Acoustical Society of America*, 107, p. 2112. doi: 10.1121/1.1354984.

Brammer, A. J. and Laroche, C. (2012) ‘Noise and communication: A three-year update’, *Noise and Health*. 14(61), pp. 281–286. doi: 10.4103/1463-1741.104894.

Brand, T. and Kollmeier, B. (2002) ‘Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests’, *The Journal of the Acoustical Society of America*, 111(6), pp. 2801–2810. doi: 10.1121/1.1479152.

Braun, D. *et al.* (1994) *Personality Profiles of US Navy Sea-Air-Land (SEAL) Personnel*, *94-8*. San Diego, California. Available at: http://oai.dtic.mil/oai/oai?verb=getRecord&amp;metadataPrefix=html&amp;identifier=ADA281692.

Bregman, A. S. (2001) ‘Auditory Scene Analysis’, in *International Encyclopedia of the Social & Behavioral Sciences*. pp. 940–942. doi: 10.1016/B0-08-043076-7/00663-X.

Brierley, C. *et al.* (2016) ‘A verified Arabic-IPA mapping for Arabic transcription technology, informed by quranic recitation, traditional Arabic linguistics, and modern phonetics’, *Journal of Semitic Studies*, 61(1), pp. 157–186. doi: 10.1093/jss/fgv035.

Bronkhorst, A. W. (2015) ‘The cocktail-party problem revisited: early processing and selection of multi-talker speech’, *Attention, Perception, & Psychophysics*. 77(5), pp. 1465–1487. doi: 10.3758/s13414-015-0882-9.

Bronkhorst, A. W. and Plomp, R. (1992) ‘Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing’, *The Journal of the Acoustical Society of America*, 92(6), pp. 3132–3139. doi: 10.1121/1.404209.

Brown, K. P., Iannelli, R. J. and Marganoff, D. P. (2017) 'Use of the Personality Assessment Inventory in fitness-for-duty evaluations of physicians', *Journal of Personality Assessment*. 99(5), pp. 465–471. doi: 10.1080/00223891.2016.1255950.

Brungart, D. S. (2001) 'Informational and energetic masking effects in the perception of two simultaneous talkers', *Journal of the Acoustical Society of America*, 109(3), pp. 1101–1111. doi: 10.1121/1.1345696.

Brungart, D. S. *et al.* (2017) 'Development and validation of the Speech Reception in Noise (SPRINT) Test', *Hearing Research*, 349, pp. 90–97. doi: 10.1016/j.heares.2017.01.008.

Brungart, D. S. *et al.* (2022) 'Assessment methods for determining small changes in hearing performance over time', *The Journal of the Acoustical Society of America*, 151(6), pp. 3866–3885. doi: 10.1121/10.0011509.

Brungart, D. S., Makashay, M. J. and Sheffield, B. M. (2021) ' Development of an 80-word clinical version of the modified rhyme test (MRT 80 ) ', *The Journal of the Acoustical Society of America*, 149(5), pp. 3311–3327. doi: 10.1121/10.0003563.

Brungart, D. S., Sheffield, B. M. and Kubli, L. R. (2014) 'Development of a test battery for evaluating speech perception in complex listening environments', *The Journal of the Acoustical Society of America*, 136(2). doi: 10.1121/1.4887440.

BSA (2018) *Recommended Procedure Pure-tone air-conduction and bone-conduction threshold audiometry with and without masking*. Bathgate. Available at: www.thebsa.org.uk (Accessed: 9 January 2019).

Burk, M. H. *et al.* (2006) 'Effect of training on word-recognition performance in noise for young normal-hearing and older hearing-impaired listeners', *Ear and Hearing*, 27(3), pp. 263–278. doi: 10.1097/01.aud.0000215980.21158.a2.

Cainer, K. E., James, C. and Rajan, R. (2008) 'Learning speech-in-noise discrimination in adult humans', *Hearing Research*, 238(1–2), pp. 155–164. doi: 10.1016/j.heares.2007.10.001.

Callister, J. D. *et al.* (1999) 'Revised NEO Personality Inventory profiles of male and female U.S. Air Force pilots', *Military Medicine*, 164(12), pp. 885–890. doi: 10.1093/milmed/164.12.885.

Campbell, J. S., Ruiz, M. A. and Moore, J. L. (2010) 'Five-factor model facet characteristics of non-aeronautically adaptable military aviators', *Aviation Space and Environmental Medicine*, 81(9), pp.

864–868. doi: 10.3357/ASEM.2761.2010.

Canzi, P. *et al.* (2016) ‘Development of a novel Italian speech-in-noise test using a roving-level adaptive method: adult population-based normative data.’, *Acta otorhinolaryngologica Italica : organo ufficiale della Societa italiana di otorinolaringologia e chirurgia cervico-facciale*, 36(6), pp. 506–512. doi: 10.14639/0392-100X-1133.

Carbonell, K. M. (2017) ‘Reliability of individual differences in degraded speech perception’, *The Journal of the Acoustical Society of America*, 142(5), pp. EL461–EL466. doi: 10.1121/1.5010148.

Carney, L. H. (2018) ‘Supra-Threshold Hearing and Fluctuation Profiles: Implications for Sensorineural and Hidden Hearing Loss’, *JARO - Journal of the Association for Research in Otolaryngology*, 19(4), pp. 331–352. doi: 10.1007/s10162-018-0669-5.

Caruso, J. C. and Cliff, N. (1997) ‘An examination of the five-factor model of normal personality variation with reliable component analysis’, *Personality and Individual Differences*, 23(2), pp. 317–325. doi: 10.1016/S0191-8869(97)00020-2.

Casali, J. G. and Robinette, M. B. (2015) ‘Effects of user training with electronically-modulated sound transmission hearing protectors and the open ear on horizontal localization ability’, *International Journal of Audiology*, 54, pp. S37–S45. doi: 10.3109/14992027.2014.973538.

Casto, K. L. and Cho, T. H. (2013) ‘In-flight speech intelligibility evaluation of a service member with sensorineural hearing loss: Case report’, *Military Medicine*, 177(9), pp. 1114–1116. doi: 10.7205/milmed-d-12-00184.

Chapman, B. P. (2007) ‘Bandwidth and fidelity on the NEO-Five Factor Inventory: Replicability and reliability of Saucier’s (1998) item cluster subcomponents’, *Journal of Personality Assessment*, 88(2), pp. 220–234. doi: 10.1080/00223890701268082.

Chen, G., Casper, W. J. and Cortina, J. M. (2001) ‘The roles of self-efficacy and task complexity in the relationships among cognitive ability, conscientiousness, and work-related performance: A meta-analytic examination’, *Human Performance*, 14(3), pp. 209–230. doi: 10.1207/S15327043HUP1403_1.

Cherry, E. C. (1953) ‘Some experiments on the recognition of speech, with one and with two ears’, *Journal of the Acoustical Society of America*, 25(5), pp. 975–979. doi: 10.1121/1.1907229.

Choi, I. *et al.* (2014) ‘Individual differences in attentional modulation of cortical responses correlate with selective attention performance’, *Hearing Research*, 314, pp. 10–19. doi: 10.1016/j.heares.2014.04.008.

Christov, F., Nelson, E. G. and Gluth, M. B. (2018) ‘Human superior olivary nucleus neuron populations in subjects with normal hearing and presbycusis’, *Annals of Otology, Rhinology and Laryngology*, 127(8), pp. 527–535. doi: 10.1177/0003489418779405.

Clasing, J. E. and Casali, J. G. (2014) ‘Warfighter auditory situation awareness: Effects of augmented hearing protection/enhancement devices and TCAPS for military ground combat applications’, *International Journal of Audiology*, 53(SUPPL.2). doi: 10.3109/14992027.2013.860489.

Coffey, E. B. J., Mogilever, N. B. and Zatorre, R. J. (2017) ‘Speech-in-noise perception in musicians: A review’, *Hearing Research*, 352, pp. 49–69. doi: 10.1016/j.heares.2017.02.006.

Colaprete, F. A. (2012) *Pre-Employment Background Investigations for Public Safety Professionals*. 1st edn. New York: CRC Press Inc. doi: https://doi.org/10.1201/b12115.

Costa, P. T., McCrae, R. and Dye, D. A. (1991) ‘Domains and facets scales for agreeableness and cconscientiousness: A revision of the NEO personality inventory’, *Journal of Personality Assessment*, 12(9), pp. 887–898.

Costa, P. T. and Mccrae, R. R. (2012) ‘The Five-Factor model, Five-Factor theory, and interpersonal psychology’, *Handbook of Interpersonal Psychology: Theory, Research, Assessment, and Therapeutic Interventions*, pp. 91–104. doi: 10.1002/9781118001868.ch6.

Cuttler, C. and Graf, P. (2007) ‘Personality predicts prospective memory task performance: An adult lifespan study’, *Scandinavian Journal of Psychology*, 48(3), pp. 215–231. doi: 10.1111/j.1467-9450.2007.00570.x.

Darr, W. (2011) ‘Military personality research: A meta-analysis of the Self Description Inventory’, *Military Psychology*, 23(3), pp. 272–296. doi: 10.1080/08995605.2011.570583.

Darr, W. and Kelloway, E. K. (2016) ‘Sifting the Big Five: Examining the criterion-related validity of facets’, *Journal of Organizational Effectiveness: People and Performance*, 3(1), pp. 2–22. doi: 10.1108/joepp-11-2015-0038.

*David Clark H3340 Headset* (2022) Available at: https://transair.co.uk/aircraft-and-airfield/airfield-equipment/ground-crew-equipment/ground-crew-headsets/david-clark-h3340-headset?code=7751 (Accessed: 5 September 2020)

Debusscher, J., Hofmans, J. and De Fruyt, F. (2017) ‘The multiple face(t)s of state conscientiousness: Predicting task performance and organizational citizenship behavior’, *Journal of Research in Personality*, 69, pp. 78–85. doi: 10.1016/j.jrp.2016.06.009.

De Raad, J., Nijhuis, F. J. N. and Willems, J. H. B. M. (2005) ‘Difference in fitness for duty among soldiers on a mission: Can these be explained by a difference in the preemployment assessment?’, *Military Medicine*, 170(9), pp. 728–734. doi: 10.7205/milmed.170.9.728.

De Raad, J. and Redekop, W. K. (2005) ‘Analysis of health factors as predictors for the functioning of military personnel: Study of the factors that predict fitness for duty and medical costs of soldiers of the Royal Netherlands Army’, *Military Medicine*, 170(1), pp. 14–20. doi: 10.7205/milmed.170.1.14.

De Sousa, K. C. *et al.* (2020) ‘Pure-tone audiometry without bone-conduction thresholds: using the digits-in-noise test to detect conductive hearing loss’, *International Journal of Audiology*, 59(10), pp. 801–808. doi: 10.1080/14992027.2020.1783585.

DeYoung, C. G., Quilty, L. C. and Peterson, J. B. (2007) ‘Between facets and domains: 10 aspects of the Big Five’, *Journal of Personality and Social Psychology*, 93(5), pp. 880–896. doi: 10.1037/0022-3514.93.5.880.

Dhanasingh, A. and Hochmair, I. (2021) ‘Thirty years of translational research behind MED-EL’, *Acta oto-laryngologica*, 141, pp. i–cxcvi. doi: 10.1080/00016489.2021.1918399.

Dobie, R. and Van Hemel, S. (2005) *Hearing Loss: Determining Eligibility for Social Security Benefits*, Washington, DC: The National Academies Press. doi: https://doi.org/10.17226/11099.

Dretsch, M. N. *et al.* (2022) ‘Variability in the stability of personality traits across a single combat deployment’, *Military Psychology*,34(4), pp. 422–431. doi: 10.1080/08995605.2021.2003147.

Dubno, J. R. (2018) ‘Beyond the audiogram: Application of models of auditory fitness for duty to assess communication in the real world’, *International Journal of Audiology*,57(5), pp. 321–322. doi: 10.1080/14992027.2018.1439677.

Dudley, N. M. *et al.* (2006) ‘A meta-analytic investigation of conscientiousness in the prediction of job performance: Examining the intercorrelations and the incremental validity of narrow traits’, *Journal of Applied Psychology*, 91(1), pp. 40–57. doi: 10.1037/0021-9010.91.1.40.

Eddins DA, Liu C. (2012) 'Psychometric properties of the coordinate response measure corpus with various types of background interference', *Journal of the Acoustical Society of America*, 131(2), pp. 177-83. doi: 10.1121/1.3678680. PMID: 22352619; PMCID: PMC3277602.

Egan, V., Deary, I. and Austin, E. (2000) ‘The NEO-FFI : Emerging British norms and an item-level analysis suggest N , A and C are more reliable than O and E’. *Personality and Individual Differences* 29 (5), 907-920. https://doi.org/10.1016/S0191-8869(99)00242-1

Elrifaey, M. A. M. *et al.* (2021) ‘Development and standardization of Arabic version of quick speech in noise test’, *Egyptian Journal of Ear, Nose, Throat and Allied Sciences*, 22(22), pp. 1–9. doi: 10.21608/EJENTAS.2021.35054.1235.

Emerson, J. W. *et al.* (2013) ‘The generalized pairs plot’, *Journal of Computational and Graphical Statistics*, 22(1), pp. 79–91. doi: 10.1080/10618600.2012.694762.

Engelmann, J. B. *et al.* (2009) ‘Combined effects of attention and motivation on visual task performance: Transient and sustained motivational effects’, *Frontiers in Human Neuroscience*, 3(MAR), pp. 1–17. doi: 10.3389/neuro.09.004.2009.

Essawy, W. M. *et al.* (2019) ‘Development and standardization of new hearing in noise test in Arabic language’, *International Journal of Otorhinolaryngology and Head and Neck Surgery*, 5(6), p. 1501. doi: 10.18203/issn.2454-5929.ijohns20194917.

Estill, C. F. *et al.* (2017) ‘Noise and neurotoxic chemical exposure relationship to workplace traumatic injuries: A review’, *Journal of Safety Research*, 60. doi: 10.1016/j.jsr.2016.11.005.

Fay, K. (2010) 'Homoscedasticity', in N.J Salkind (ed.) *Encyclopedia of Research Design* Los Angeles, CA: Sage Reference USA, pp.580-583.

Ferguson, M. and Henshaw, H. (2015) ‘How does auditory training work? Joined-up thinking and listening’, *Seminars in Hearing*, 36(4), pp. 237–249. doi: 10.1055/s-0035-1564456.

Field, A. P. (2009) *Discovering statistics using SPSS : (and sex and drugs and rock ‘n’ roll)*. 3rd edn. SAGE Publications.

Fleming, K. A., Heintzelman, S. J. and Bartholow, B. D. (2016) ‘Specifying associations between conscientiousness and executive functioning: Mental set shifting, not prepotent response inhibition or working memory updating’, *Journal of Personality*, 84(3), pp. 348–360. doi: 10.1111/jopy.12163.

Frost, J. (2019) *Standard Error of the Regression - Statistics By Jim*. Available at: https://statisticsbyjim.com/regression/standard-error-regression-vs-r-squared/ (Accessed: 28 November 2019).

Fründ, I., Haenel, N. V. and Wichmann, F. A. (2011) ‘Inference for psychometric functions in the presence of nonstationary behavior’, *Journal of Vision*, 11(6), pp. 1–19. doi: 10.1167/11.6.16.

Füllgrabe, C. and Rosen, S. (2016) ‘On the (un)importance of working memory in speech-in-noise processing for listeners with normal hearing thresholds’, *Frontiers in Psychology*. 7:1268. doi:

10.3389/fpsyg.2016.01268.

Gallun, F. J. *et al.* (2012) 'Performance on tests of central auditory processing by individuals exposed to high-intensity blasts', *Journal of Rehabilitation Research and Development*, 49(7), pp. 1005–1024. doi: 10.1682/JRRD.2012.03.0038.

General Organisation for Retirement (2004) *Regulations of the General Organisation of Retirement*. Kingdom of Saudi Arabia: National Center for documentation and Archives.

Giavarina, D. (2015) 'Understanding Bland Altman analysis', *Biochemia Medica*, 25(2), pp. 141–151. doi: 10.11613/BM.2015.015.

Gieseler, A. *et al.* (2017) 'Auditory and non-auditory contributions for unaided speech recognition in noise as a function of hearing aid use', *Frontiers in Psychology*, 8(219). https://doi.org/10.3389/fpsyg.2017.00219

Giguère, C. *et al.* (2008) 'Functionally-based screening criteria for hearing-critical jobs based on the Hearing in Noise Test.', *International journal of audiology*, 47(6), pp. 319–28. doi: 10.1080/14992020801894824.

Giguѐre, C. *et al.* (2019) 'Development of hearing standards for Ontario's Constable Selection System', *International Journal of Audiology*, 58(11), pp. 798–804. doi: 10.1080/14992027.2019.1617438.

Giguѐre, C. and Berger, E. H. (2016) 'Speech recognition in noise under hearing protection: A computational study of the combined effects of hearing loss and hearing protector attenuation', *International Journal of Audiology*, 55(December 2015), pp. S30–S40. doi: 10.3109/14992027.2015.1129460.

Glyde, H. *et al.* (2011) 'Problems hearing in noise in older adults: A review of spatial processing disorder', *Trends in Amplification*, 15(3), pp. 116–126. doi: 10.1177/1084713811424885.

Gnansia, D. *et al.* (2009) 'Effects of spectral smearing and temporal fine structure degradation on speech masking release', *The Journal of the Acoustical Society of America*, 125(6), pp. 4023–4033. doi: 10.1121/1.3126344.

Gogniat, M. A. *et al.* (2022) 'Differential item functioning: An examination of the NEO-FFI by sex in older adults', *SAGE Open*, 12(1). doi: 10.1177/21582440221086607.

Goossens, T. *et al.* (2017) 'Masked speech perception across the adult lifespan: Impact of age and hearing impairment', *Hearing Research*, pp. 109–124. doi: 10.1016/J.HEARES.2016.11.004.

Gordon-Salant, S. and Cole, S. S. (2016) 'Effects of age and working memory capacity on speech recognition performance in noise among listeners with normal hearing.', *Ear and hearing*, 37(5), pp. 593–602. doi: 10.1097/AUD.0000000000000316.

Grant, K. W. *et al.* (2021) 'Estimated prevalence of functional hearing difficulties in blast-exposed service members with normal to near-normal-hearing thresholds', *Ear and Hearing*, pp. 1615–1626. doi: 10.1097/AUD.0000000000001067.

Grantham, M. A. M. (2012) 'Noise-Induced Hearing Loss and Tinnitus: Challenges for the Military', in. Springer, New York, NY, pp. 27–38. doi: 10.1007/978-1-4419-9523-0_3.

Griffin, B. and Hesketh, B. (2003) 'Adaptable Behaviours for Successful Workand Career Adjustment', *Australian Journal of Psychology*, 55(2), pp. 65–73.

Haahr, M. (2022) *Introduction to Randomness and Random Numbers*. Available at: https://www.random.org/randomness/ (Accessed: 04 September 2019).

Haboosheh, R. and Brown, S. (2012) 'Workplace hearing loss', *British Columbia Medical Journal*, 54(4), p. 175.

Hackett, R. D. (2002) 'Understanding and predicting work performance in the Canadian military', *Canadian Journal of Behavioural Science*, 34(2), pp. 131–140. doi: 10.1037/h0087163.

Hagerman, B. (1997) 'Attempts to develop an efficient speech test in fully modulated noise', *Scandinavian Audiology*, 26(2), pp. 93–98. doi: 10.3109/01050399709074980.

Hagerman, B. and Kinnefors, C. (1995) 'Efficient adaptive methods for measuring speech reception threshold in quiet and in noise', *Scandinavian Audiology*, 24(1), pp. 71–77. doi: 10.3109/01050399509042213.

Harari, M. B., Naemi, B. and Viswesvaran, C. (2019) 'Is the validity of conscientiousness stable across time? Testing the role of trait bandwidth', *Journal of Occupational and Organizational Psychology*, 92(1), pp. 212–220. doi: 10.1111/joop.12241.

Hassan, S., Akhtar, N. and Yılmaz, A. K. (2016) 'Impact of the conscientiousness as personality trait on both job and organizational performance.', *Journal of Managerial Sciences*, 10(1), pp. 1–14.

Hasson, D. *et al.* (2011) 'Stress and prevalence of hearing problems in the Swedish working population', *BMC Public Health*, 11(130), pp. 1–12.

*Hawk MIM-23 low medium altitude ground to air missile technical data sheet specifications* (2012) *Army Recognition*. Available at:

https://www.armyrecognition.com/united_states_american_missile_system_vehicle_uk/hawk_mim-23_low_medium_altitude_ground_to_air_missile_technical_data_sheet_specifications_pictures.html (Accessed: 29 December 2019).

Hearing Link (2018) *Facts about deafness &amp; hearing loss - Hearing Link*, *Hearing Link*. Available at: https://www.hearinglink.org/your-hearing/about-hearing/facts-about-deafness-hearing-loss/ (Accessed: 27 October 2019).

Herzberg, P. Y. and Brähler, E. (2006) 'Assessing the Big-Five personality domains via short forms a cautionary note and a proposal', *European Journal of Psychological Assessment*, 22(3), pp. 139–148. doi: 10.1027/1015-5759.22.3.139.

Holmes, E. and Griffiths, T. D. (2019) 'Hearing thresholds and fundamental auditory grouping processes predict difficulties with speech-in-noise perception', *Scientific Reports*. Springer Science and Business Media LLC, 9(1).

Hoover, E. C., Souza, P. E. and Gallun, F. J. (2017) 'Auditory and cognitive factors associated with speech-in-noise complaints following mild traumatic brain injury', *Journal of the American Academy of Audiology*, 28(4), pp. 325–339. doi: 10.3766/jaaa.16051.

Hopkins, K. and Moore, B. C. J. (2010) 'The importance of temporal fine structure information in speech at different spectral regions for normal-hearing and hearing-impaired subjects', *The Journal of the Acoustical Society of America*, 127(3), pp. 1595–1608. doi: 10.1121/1.3293003.

Houben, R. *et al.* (2014) 'Development of a Dutch matrix sentence test to assess speech intelligibility in noise', *International Journal of Audiology*, 53(10), pp. 760–763. doi: 10.3109/14992027.2014.920111.

Howell, D.C. (2010) 'Intraclass correlation', in N.J Salkind (ed.) *Encyclopedia of Research Design* Los Angeles, CA: Sage Reference USA, pp.636-640.

Huddle, M. G. *et al.* (2017) 'The Economic Impact of Adult Hearing Loss', *JAMA Otolaryngology–Head & Neck Surgery*, 143(10), p. 1040. doi: 10.1001/jamaoto.2017.1243.

Huijzer, R. *et al.* (2022) 'Personality traits of Special Forces Operators: Comparing Commandos, candidates, and controls', *Sport, Exercise, and Performance Psychology*, 11(3), pp. 369–381. doi: 10.1037/spy0000296.

Hull, D. M. *et al.* (2010) 'An item-level examination of the factorial validity of NEO Five-Factor Inventory scores', *Educational and Psychological Measurement*, pp. 1021–1041. doi:

10.1177/0013164410378091.

Humes, L. E. (2007) ‘Speech to Older Adults’, *Journal of American Academy of Audiology*, 603, pp. 590–603.

Hurtz, G. M. and Donovan, J. J. (2000) ‘Personality and job performance: The Big Five revisited’, *Journal of Applied Psychology*, 85(6), pp. 869–879. doi: 10.1037//002l-9010.85.6.869.

Hwang, J. S., Kim, K. H. and Lee, J. H. (2017) ‘Factors affecting sentence-in-noise recognition for normal hearing listeners and listeners with hearing loss.’, *Journal of audiology & otology*. Korean Audiological Society, 21(2), pp. 81–87. doi: 10.7874/jao.2017.21.2.81.

Igo, R.P.Jr. (2010) 'Influential data points', in N.J Salkind (ed.) *Encyclopedia of Research Design* Los Angeles, CA: Sage Reference USA, p.601.

Ihlefeld, A. and Shinn-Cunningham, B. (2008) ‘Disentangling the effects of spatial cues on selection and formation of auditory objects.’, *The Journal of the Acoustical Society of America*, 124(4), pp. 2224–35. doi: 10.1121/1.2973185.

Imhof, M. and Spaeth-Hilbert, T. (2013) ‘The role of motivation, cognition, and conscientiousness for academic achievement’, *International Journal of Higher Education*, 2(3), pp. 69–80. doi: 10.5430/ijhe.v2n3p69.

Innes-Brown, H. *et al.* (2016) ‘Towards Objective Measures of Functional Hearing Abilities’, in. Springer, Cham, pp. 315–325. doi: 10.1007/978-3-319-25474-6_33.

Jackson, J. J. *et al.* (2009) ‘Not all conscientiousness scales change alike: A multimethod, multisample study of age differences in the facets of conscientiousness’, *Journal of Personality and Social Psychology*, 96(2), pp. 446–459. doi: 10.1037/a0014156.Not.

Jansen, S. *et al.* (2013) ‘Efficient hearing screening in noise-exposed listeners using the digit triplet test’, *Ear and Hearing*, 34(6), pp. 773–778. doi: 10.1097/AUD.0b013e318297920b.

Jansen, S. *et al.* (2014) ‘Exploring the sensitivity of speech-in-noise tests for noise-induced hearing loss’, *International Journal of Audiology*, 53(3), pp. 199–205. doi: 10.3109/14992027.2013.849361.

Jiang, C., Wang, D. and Zhou, F. (2009) ‘Personality traits and job performance in local government organizations in China’, *Social Behavior and Personality*, 37(4), pp. 451–458. doi: 10.2224/sbp.2009.37.4.451.

Jiang, G., Garris, C. P. and Aldamer, S. (2018) ‘Individualism behind collectivism: A reflection from

Saudi volunteers', *Voluntas*, 29(1), pp. 144–159. doi: 10.1007/s11266-017-9872-y.

Johansen, R. B., Laberg, J. C. and Martinussen, M. (2014) 'Military identity as predictor of perceived military competence and skills', *Armed Forces and Society*, 40(3), pp. 521–543. doi: 10.1177/0095327X13478405.

Judge, T. A. *et al.* (2008) 'The contributions of personality to organizational behavior and psychology : Findings , criticisms , and future research directions', *Social and Personality Psychology Compass,*2, pp. 1982–2000.

Keidser, G. *et al.* (2015) 'Cognitive spare capacity: evaluation data and its association with comprehension of dynamic conversations', *Frontiers in Psychology*, 6, p. 597. doi: 10.3389/fpsyg.2015.00597.

Kelly, M. P., Mulligan, K. P. and Monohan, M. C. (2010) 'Fitness for Duty', in *Military Neuropsychology*. New York: Springer, pp. 57–80. doi: 10.1007/978-3-319-93497-6_17.

Killion, M. C. *et al.* (2004) 'Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners', *The Journal of the Acoustical Society of America*, 116(4), pp. 2395–2405. doi: 10.1121/1.1784440.

Kingdom, A. A. F. and Prins, N. (2010) *Psychophysics A Practical Introduction*. Elsevier Academic Press

Koifman, S. *et al.* (2016) 'Comparing evaluation data of the digit triplet test for Arabic , Hebrew and Persian', 19 Jahrestagung der Deutschen Gesellschaft für Audiologie, March. Available at https://www.researchgate.net/publication/306129165_Comparing_evaluation_data_of_the_digit_triplet_test_for_Arabic_Hebrew_and_Persian (Accessed 1 February 2019)

Kollmeier, B. *et al.* (2015) 'The multilingual matrix test: Principles, applications, and comparison across languages: A review', *International Journal of Audiology*, 54, pp. 3–16. doi: 10.3109/14992027.2015.1020971.

Koo, T. K. and Li, M. Y. (2016) 'A guideline of selecting and reporting intraclass correlation coefficients for reliability research', *Journal of Chiropractic Medicine*, 15(2), pp. 155–163. doi: 10.1016/j.jcm.2016.02.012.

Kooser, C. (2013) 'Hearing loss and employment in the United States', *Work*, 46(2), pp. 181–186. doi: 10.3233/WOR-131746.

Körner, A. *et al.* (2015) 'Efficient and valid assessment of personality traits: Population norms of a brief version of the NEO Five-Factor Inventory (NEO-FFI)', *Archives of Psychiatry and Psychotherapy*, 17(1), pp. 21–32. doi: 10.12740/APP/36086.

Kramer, S. E., Kapteyn, T. S. and Houtgast, T. (2006) 'Occupational performance: Comparing normally-hearing and hearing-impaired employees using the Amsterdam Checklist for Hearing and Work', *International Journal of Audiology*, 45(9), pp. 503–512. doi: 10.1080/14992020600754583.

Kurtz, J.E. (2020). NEO Inventories. In: Zeigler-Hill, V., Shackelford, T.K. (eds) Encyclopedia of Personality and Individual Differences. Springer, Cham. https://doi.org/10.1007/978-3-319-24612-3_940

Lakens, D. (2013) 'Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs', *Frontiers in Psychology*, 4(NOV), pp. 1–12. doi: 10.3389/fpsyg.2013.00863.

Laroche, C. *et al.* (2003) 'An approach to the development of hearing standards for hearing-critical jobs.', *Noise & health*, 6(21), pp. 17–37. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14965451 (Accessed: 22 May 2018).

Laroche, C. *et al.* (2011), ' Update on fitness standards for hearing-critical jobs'*, "10th International Congress on Noise as a Public Health Problem (ICBEN),"* in *Proceedings of the Institute of Acoustics* . (Pt.3). London, pp. 234–241. Available at: https://www.researchgate.net/profile/Christian_Giguere2/publication/283556245_Update_on_fitness_standards_for_hearing-critical_jobs/links/563ec37d08ae45b5d28c69fa.pdf (Accessed: January 2, 2019).

Leclercq, F., Renard, C. and Vincent, C. (2018) 'Speech audiometry in noise: Development of the French-language VRB (vocale rapide dans le bruit) test', *European Annals of Otorhinolaryngology, Head and Neck Diseases*, 135(5), pp. 315–319. doi: 10.1016/J.ANORL.2018.07.002.

Lee, J. E. C., Sudom, K. A. and Zamorski, M. A. (2013) 'Longitudinal analysis of psychological resilience and mental health in Canadian military personnel returning from overseas deployment', *Journal of Occupational Health Psychology*, 18(3), pp. 327–337. doi: 10.1037/a0033059.

Leek, M. R. (2001) 'Adaptive procedures in psychophysical research', *Perception & Psychophysics*, 63(8), pp. 1279–1292. Available at: papers2://publication/uuid/D1338FF5-084B-427E-9C17-CAF072C75B02.

Lie, A. *et al.* (2016) 'Occupational noise exposure and hearing: a systematic review', *International*

*Archives of Occupational and Environmental Health*, 89(3), pp. 351–372. doi: 10.1007/s00420-015-1083-5.

Lin, J. *et al.* (2019) ‘Overload and automation-dependence in a multi-UAS simulation: Task demand and individual difference factors’, *Journal of Experimental Psychology: Applied*, 26(2), pp. 218–235. doi: 10.1037/xap0000248.

Lin, L. and Xu, C. (2020) ‘Arcsine-based transformations for meta-analysis of proportions: Pros, cons, and alternatives’, *Health Science Reports*, 3(3), pp. 1–6. doi: 10.1002/hsr2.178.

Lüdtke, O., Trautwein, U. and Husemann, N. (2009) ‘Goal and personality trait development in a transitional period: Assessing change and stability in personality development’, *Personality and Social Psychology Bulletin*, 35(4), pp. 428–441. doi: 10.1177/0146167208329215.

Macpherson, A. and Akeroyd, M. A. (2014) ‘Variations in the slope of the psychometric functions for speech intelligibility: A systematic survey’, *Trends in Hearing*, 18, pp. 10–16. doi: 10.1177/2331216514537722.

Magalhães, E. *et al.* (2014) ‘NEO-FFI: Psychometric properties of a short personality inventory in Portuguese context’, *Psicologia: Reflexao e Critica*, 27(4), pp. 642–657. doi: 10.1590/1678-7153.201427405.

Mandrekar, J. N. (2010) ‘Receiver operating characteristic curve in diagnostic test assessment’, *Journal of Thoracic Oncology*. International Association for the Study of Lung Cancer, 5(9), pp. 1315–1316. doi: 10.1097/JTO.0b013e3181ec173d.

Manga, D., Ramos, F., Morán, C. (2004) ‘The Spanish norms of the NEO Five-Factor Inventory: New data and analyses for its improvement’, *International Journal of Psychology and Psychological Therapy*, 4(3), pp. 639–648.

Maruthy, S., Kumar, U. A. and Gnanateja, G. N. (2017) ‘Functional interplay between the putative measures of rostral and caudal efferent regulation of speech perception in noise’, *JARO - Journal of the Association for Research in Otolaryngology*, 18(4), pp. 635–648. doi: 10.1007/s10162-017-0623-y.

Masterson, E. A. *et al.* (2016) ‘Hearing impairment among noise-exposed workers - United States, 2003-2012’, *MMWR.Morbidity and mortality weekly report*, 65(15), p. 389. Available at: http://sdl.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwnV1LT8JAEJ4oHjQxPvCFaLInT1boLt1tD8aAYnyFGMAz2W23CQkCUo3-fGe6RYgHD956adLZmZ1-880LQPCLuvfLJ6RCS259JaIGRhy-VakKBQZ1XASmzk3eBP8Q9bphrxU8Fk39xBQ4bcdZO65k0lMpHnNp0HoUV1wdTV982iNFKVbi50aq7AmfK7wpq612p3.

Mathias, C.W. (2010) ‘Sensitivity’, in N.J Salkind (ed.) *Encyclopedia of Research Design* Los Angeles, CA: Sage Reference USA, PP.1337-1338.

Mattys, S. L. *et al.* (2012) ‘Speech recognition in adverse conditions: A review’, *Language and Cognitive Processes*, 27(7–8), pp. 953–978. doi: 10.1080/01690965.2012.705006.

McCormack, L. and Mellor, D. (2002) ‘The role of personality in leadership: An application of the Five-Factor model in the Australian military’, *Military Psychology*, 14(3), pp. 179–197. doi: 10.1207/S15327876MP1403_1.

McCrae, R. R. *et al.* (1999) ‘Age differences in personality across the adult life span: Parallels in five cultures.’, *Developmental psychology*, 35(2), pp. 466–477. doi: 10.1037/0012-1649.35.2.466.

McCrae, R. R. *et al.* (2004) ‘Age differences in personality traits across cultures: Self-report and observer perspectives’, *European Journal of Personality*, 18(2), pp. 143–157. doi: 10.1002/per.510.

McDermott, J. H. (2009) ‘The cocktail party problem.’, *Current biology : CB*, 19(22), pp. R1024–R1027. doi: 10.1016/j.cub.2009.09.005.

McLaughlin, D. J. *et al.* (2018) ‘Coping with adversity: Individual differences in the perception of noisy and accented speech’, *Attention, Perception, and Psychophysics*. 80(6), pp. 1559–1570. doi: 10.3758/s13414-018-1537-4.

McNeil, J. A. and Morgan, C. A. (2010) *Cognition and decision making in extreme environments*, *Military Neuropsychology*. Edited by C. H. Kennedy and J. L. Moore. New York, NY: Springer.

Merten, N. *et al.* (2022) ‘The associations of hearing sensitivity and different cognitive functions with perception of speech-in-noise’, *Ear and Hearing*, 43(3), pp. 984–992.

Meyer, R. D., Dalal, R. S. and Bonaccio, S. (2009) ‘A meta-analytic investigation into the moderating effects of situational strength on the conscientiousness–performance relationship’, *Journal of Organizational Behavior*, 30, pp. 1077–1102. doi: 10.1002/job.602.

Meyer, R. D., Dalal, R. S. and Hermida, R. (2010) ‘A Review and synthesis of situational strength in the organizational sciences’, *Journal of Management*, 36(1), pp. 121–140. doi: 10.1177/0149206309349309.

Miller, N. L., Matsangas, P. and Shattuck, L. G. (2008) ‘Fatigue and its effect on performance in military environments’, *Performance Under Stress*, pp. 231–249.

Minbashian, A., Wood, R. E. and Beckmann, N. (2010) ‘Task-contingent conscientiousness as a unit of personality at work’, *Journal of Applied Psychology*, 95(5), pp. 793–806. doi: 10.1037/a0020016.

Mishra, S. K., Saxena, U. and Rodrigo, H. (2022) ‘Extended high-frequency hearing impairment despite a normal audiogram: Relation to early aging, speech-in-noise perception, cochlear function, and routine earphone use’, *Ear and Hearing*, 43(3), pp. 822–835. doi: 10.1097/AUD.0000000000001140.

Mohaisen (2013) ‘Factorial structure of the Big Five Personality Factors Inventory among Palestinian university students in Gaza’, *Journal of Educational and Psychological Sciences-Bahrain*, 3 (14), pp. 387–416. Available at: http://search.mandumah.com/Record/466524%0A.

*MOH Covid19 Awareness* (2021) Available at: https://covid19awareness.sa (Accessed: 27 October 2021).

Moldzio, T. *et al.* (2021) ‘Differentiated measurement of conscientiousness and emotional stability in an occupational context–greater effort or greater benefit?’, *European Journal of Work and Organizational Psychology*, 30(2), pp. 192–205. doi: 10.1080/1359432X.2020.1866066.

Moore, B. C. . (2013) *An Introduction to the Psychology of Hearing*. sixth. Leiden, The Netherlands: Brill. available from: <https://brill.com/view/title/24210> [Accessed 08 December 2022]

Moore, B. C. J. (2008) ‘The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people’, *JARO - Journal of the Association for Research in Otolaryngology*, 9(4), pp. 399–406. doi: 10.1007/s10162-008-0143-x.

Moore, T. J. (1981) ‘Aural Communication in Aviation’, in *Voice Communications Jamming Research*, pp. 2:1-2:6.

Morey, R. D. *et al.* (2016) ‘The fallacy of placing confidence in confidence intervals’, *Psychonomic Bulletin and Review*, 23(1), pp. 103–123. doi: 10.3758/s13423-015-0947-8.

Motlagh Zadeh, L. *et al.* (2019) ‘Extended high-frequency hearing enhances speech perception in noise’, *Proceedings of the National Academy of Sciences*, 116(47), pp. 23753–23759.

Muralidhar, K. (2003) *Monte carlo simulation*, *Elsevier*. Edited by H. Bidgoli. doi: 10.1016/B0-12-227240-4/00114-3.

Murman DL. (2015) 'The Impact of Age on Cognition', *Seminars in Hearing*, 36(3), pp. 111-21. doi: 10.1055/s-0035-1555115. PMID: 27516712; PMCID: PMC4906299.

Murray, G. *et al.* (2003) ‘Neo Five-Factor Inventory scores: Psychometric properties in a community sample’, *Measurement and Evaluation in Counseling and Development*, 36(3), pp. 140–149. doi: 10.1080/07481756.2003.11909738.

Musiek, F. E. *et al.* (2017) ‘Perspectives on the pure-tone audiogram’, *Journal of the American Academy of Audiology*, 28(7), pp. 655–671. doi: 10.3766/jaaa.16061.

National Academies of Sciences and Medicine., E. and (2019) *Functional Assessment for Adults with Disabilities*, *Functional Assessment for Adults with Disabilities*. Edited by P. A. Volberding, C. M. Spicer, and J. L. Flaubert. Washington, DC: The National Academies Press. doi: 10.17226/25376.

Neyman, J. (1937) ‘Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability’, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 236(767), pp. 333–380. doi: 10.1098/rsta.1937.0005.

Ng, K. Y., Ang, S. and Chan, K. Y. (2008) ‘Personality and leader effectiveness: A moderated mediation model of leadership self-efficacy, job demands, and job autonomy’, *Journal of Applied Psychology*, 93(4), pp. 733–743. doi: 10.1037/0021-9010.93.4.733.

Nielsen, J. B. and Dau, T. (2009) ‘Development of a Danish speech intelligibility test’, *International Journal of Audiology*, 48(10), pp. 729–741. doi: 10.1080/14992020903019312.

Nilsson, M., Soli, S. D. and 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), pp. 1085–1099. doi: 10.1121/1.408469.

Occupational Safety and Health Administration (2002) ‘Hearing conservation.’, *Hearing Conservation*. doi: 10.1177/014107687807101119.

Ozimek, E. *et al.* (2009) ‘Polish sentence tests for measuring the intelligibility of speech in interfering noise’, *International Journal of Audiology*, 48, pp. 433–443. doi: 10.1080/14992020902725521.

Paglialonga, A. *et al.* (2020) ‘An automated speech-in-noise test for remote testing: Development and preliminary evaluation’, *American Journal of Audiology*, 29(3 Special Issue), pp. 564–576. doi: 10.1044/2020_AJA-19-00071.

Paulus, M., Hazan, V. and Adank, P. (2020) ‘The relationship between talker acoustics, intelligibility, and effort in degraded listening conditions’, *The Journal of the Acoustical Society of America*, 147(5), pp. 3348–3359. doi: 10.1121/10.0001212.

Payne, W. and Harvey, J. (2010) ‘A framework for the design and development of physical employment tests and standards’, *Ergonomics*, 53(7), pp. 858–871. doi: 10.1080/00140139.2010.489964.

Pedersen, E. R. and Juhl, P. M. (2017) ‘Simulated critical differences for speech reception thresholds’, *Journal of Speech, Language, and Hearing Research*, 60, pp. 238–250.

Peele, J. (2018) OSF/*Peelle lab figures/Human auditory pathway*. Available at: https://osf.io/u2gxc/ (Accessed: 4 December 2022).

Peelle, J. E. and Wingfield, A. (2005) ‘Dissociations in perceptual learning revealed by adult age differences in adaptation to time-compressed speech’, *Journal of Experimental Psychology: Human Perception and Performance*, 31(6), pp. 1315–1330. doi: 10.1037/0096-1523.31.6.1315.

Perry, S. J. *et al.* (2010) ‘P = f (conscientiousness × ability): Examining the facets of conscientiousness’, *Human Performance*, 23(4), pp. 343–360. doi:10.1080/08959285.2010.501045.

Phatak, S. A. *et al.* (2019) ‘Clinical assessment of functional hearing deficits: Speech-in-Noise performance’, *Ear and Hearing*, 40(2), pp. 426–436. doi: 10.1097/AUD.0000000000000635.

Picard, M. *et al.* (1999) ‘Speech audiometry in noise-exposed workers: The SRT-PTA relationship revisited’, *Audiology*, 38(1), pp. 30–43. doi: 10.3109/00206099909073000.

Pichora-Fuller, M. K., Schneider, B. A. and Daneman, M. (1995) ‘How young and old adults listen to and remember speech in noise.’, *The Journal of the Acoustical Society of America*, 97(1), pp. 593–608. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7860836 (Accessed: 5 June 2018).

Picou, E. M., Gordon, J. and Ricketts, T. A. (2016) ‘The effects of noise and reverberation on listening effort in adults with normal hearing’, *Ear and Hearing*, 37(1), pp. 1–13. doi: 10.1097/AUD.0000000000000222.

Pienkowski, M. (2017) ‘On the etiology of listening difficulties in noise despite clinically normal audiograms’, *Ear and Hearing*, 38(2). doi: 10.1097/AUD.0000000000000388.

Plomp, R. (1986) ‘A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired.’, *Journal of speech and hearing research*, 29(2), pp. 146–54. Available at: http://www.ncbi.nlm.nih.gov/pubmed/3724108 (Accessed: 3 January 2019).

Polspoel, S. *et al.* (2022) ‘The importance of extended high-frequency speech information in the recognition of digits, words, and sentences in quiet and noise’, *Ear and Hearing*, 43(3), pp. 913–

920. doi: 10.1097/AUD.0000000000001142.

Pürner, D., Schirkonyer, V. and Janssen, T. (2022) “Changes in the peripheral and central auditory performance in the elderly—a cross-sectional study,” *Journal of Neuroscience Research*, 100(9), pp. 1791–1811. Available at: https://doi.org/10.1002/jnr.25068.

Le Prell, C. G. (2019) ‘Effects of noise exposure on auditory brainstem response and speech-in-noise tasks: a review of the literature’, *International Journal of Audiology*, 58(sup1), pp. S3–S32. doi: 10.1080/14992027.2018.1534010.

Le Prell, C. G. and Clavier, O. H. (2017) ‘Effects of noise on speech recognition: Challenges for communication by service members’, *Hearing Research*, pp. 76–89. doi: 10.1016/j.heares.2016.10.004.

Pulakka, H. *et al.* (2012) ‘Conversational quality evaluation of artificial bandwidth extension of telephone speech’, *The Journal of the Acoustical Society of America*, 132(2), pp. 848–861. doi: 10.1121/1.4730882.

Pulakos, E. D. *et al.* (2000) ‘Adaptability in the workplace: Development of a taxonomy of adaptive performance’, *Journal of Applied Psychology*, 85(4), pp. 612–624. doi: 10.1037/0021-9010.85.4.612.

Punch, R. (2016) ‘Employment and adults who are deaf or hard of hearing: Current status and experiences of barriers, accomodations, and stress in the workplace’, *American Annals of the Deaf*, 161(3), pp. 384–397.

Puvvada, K. C. and Simon, J. Z. (2017) ‘Cortical representations of speech in a multitalker auditory scene’, *Journal of Neuroscience*, 37(38), pp. 9189–9196. doi: 10.1523/JNEUROSCI.0938-17.2017.

Rhebergen, K. S., Versfeld, N. J. and Dreschler, W. A. (2008) ‘Learning effect observed for the speech reception threshold in interrupted noise with normal hearing listeners’, *International Journal of Audiology*, 47(4), pp. 185–188. doi: 10.1080/14992020701883224.

Richardson, M. and Abraham, C. (2009) ‘Conscientiousness and achievement motivation predict performance’, *European Journal of Personality*, 23, pp. 589–605. doi: 10.1002/per.732.

Robson, S. *et al.* (2017) Fit for Duty? Evaluating the Physical Fitness Requirements of Battlefield Airmen, *Rand Health Quarterly,* 7(2), p.8 doi: 10.7249/rr618.

Rolland, J. P., Parker, W. D. and Stumpf, H. (1998) ‘A psychometric examination of the French translations of the NEO-PI-R and NEO-FFI’, *Journal of Personality Assessment*, 71(2), pp. 269–291.

Rönnberg, J., Holmer, E. and Rudner, M. (2019) ‘Cognitive hearing science and ease of language understanding’, *International Journal of Audiology*, 58(5), pp. 247–261. doi: 10.1080/14992027.2018.1551631.

Ross, B., Dobri, S. and Schumann, A. (2020) ‘Speech-in-noise understanding in older age: The role of inhibitory cortical responses’, *European Journal of Neuroscience*, 51(3), pp. 891–908. doi: 10.1111/ejn.14573.

Roup, C. M. *et al.* (2011) ‘Evaluation of a telephone speech-enhancement algorithm among older adults with hearing loss’, *Journal of Speech, Language, and Hearing Research*, 54(5), pp. 1477–1483. doi: 10.1044/1092-4388(2011/10-0181).

Ruggles, D., Bharadwaj, H. and Shinn-Cunningham, B. G. (2011) ‘Normal hearing is not enough to guarantee robust encoding of suprathreshold features important in everyday communication’, *Proceedings of the National Academy of Sciences of the United States of America*, 108(37), pp. 15516–15521. doi: 10.1073/pnas.1108912108.

Sackett, P. R. *et al.* (2017) ‘Supplemental material for individual differences and their measurement: A review of 100 years of research’, *Journal of Applied Psychology*, 102(3), pp. 254–273. doi: 10.1037/apl0000151.supp.

Saiz-Alía, M., Forte, A. E. and Reichenbach, T. (2019) ‘Individual differences in the attentional modulation of the human auditory brainstem response to speech inform on speech-in-noise deficits’, *Scientific Reports*, 9(1), pp. 1–10. doi: 10.1038/s41598-019-50773-1.

Salgado, J. F., Moscoso, S. and Berges, A. (2013) ‘Conscientiousness, its facets, and the prediction of job performance ratings: Evidence against the narrow measures’, *International Journal of Selection and Assessment*, 21(1), pp. 74–84. doi: 10.1111/ijsa.12018.

Sanchez-Lopez, R. *et al.* (2021) ‘Auditory tests for characterizing hearing deficits in listeners with various hearing abilities: The BEAR Test Battery’, *Frontiers in Neuroscience*, 15(September), pp. 1–19. doi: 10.3389/fnins.2021.724007.

Sassenberg, T. A. *et al.* (2023) 'Conscientiousness associated with efficiency of the salience/ventral attention network: Replication in three samples using individualized parcellation', *Neuroimage,*272:120081, doi: 10.1101/2022.06.07.495168.

Sataloff, R. T. and Sataloff, J. (2006) *Occupational Hearing Loss, Third Edition*. in R. T. Sataloff and J. Sataloff (ed.). Florida: CRC Press, Taylor and Francis Group.

Saunders, G. H. and Haggard, M. P. (1992) ‘The clinical assessment of “obscure auditory dysfunction” (OAD) 2. Case control analysis of determining factors’, *Ear and Hearing*, 13(4), pp. 241–254. doi: 10.1097/00003446-199208000-00006.

Schaafsma, F. *et al.* (2016) ‘Pre-employment examinations for preventing injury, disease and sick leave in workers (Review)’, *Cochrane Database of Systematic Reviews*, (1). doi: 10.1002/14651858.CD008881.pub2.www.cochranelibrary.com.

Schlueter, A. *et al.* (2016) ‘Normal and Time-Compressed Speech’, *Trends in Hearing*, 20, pp. 1–13. doi: 10.1177/2331216516669889.

Schoof, T. and Rosen, S. (2014) ‘The role of auditory and cognitive factors in understanding speech in noise by normal-hearing older listeners’, *Frontiers in Aging Neuroscience*, 6, p. 307. doi: 10.3389/fnagi.2014.00307.

Semeraro, H. *et al.* (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. doi: 10.4103/1463-1741.153401.

Semeraro, H. D. (2015) *Developing a measure of auditory fitness for duty for military personnel*. University of Southampton. Available at: https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.680720? (Accessed: 2 January 2019).

Semeraro, H. D. *et al.* (2017) ‘Development and evaluation of the British English coordinate response measure speech-in-noise test as an occupational hearing assessment tool’, *International Journal of Audiology*, 56(10), pp. 749–758. doi: 10.1080/14992027.2017.1317370.

Serra, C. *et al.* (2007) ‘Criteria and methods used for the assessment of fitness for work: A systematic review’, *Occupational and Environmental Medicine*, 64(5), pp. 304–312. doi: 10.1136/oem.2006.029397.

Shaffer, J. A. and Postlethwaite, B. E. (2013) ‘The validity of conscientiousness for predicting job performance: A meta-analytic test of two hypotheses’, *International Journal of Selection and Assessment*, 21(2), pp. 183–199. doi: 10.1111/ijsa.12028.

Sharma, S., Tripathy, R. and Saxena, U. (2017) ‘Critical appraisal of speech in noise tests: a systematic review and survey’, *International Journal of Research in Medical Sciences*, 5(1), pp. 13–21. doi: http://dx.doi.org/10.18203/2320-6012.ijrms20164525.

Sheffield, B. *et al.* (2015) ‘The relationship between hearing acuity and operational performance in dismounted combat’, *Proceedings of the Human Factors and Ergonomics Society*, 2015-January, pp. 1346–1350. doi: 10.1177/1541931215591223.

Sheikh Rashid, M. and Dreschler, W. A. (2018) ‘Accuracy of an internet-based speech-in-noise hearing screening test for high-frequency hearing loss: incorporating automatic conditional rescreening’, *International Archives of Occupational and Environmental Health*, 91(7), pp. 877–885. doi: 10.1007/s00420-018-1332-5.

Shinn-Cunningham, B., Best, V. and Lee, A. K. C. (2017) *The Auditory System at the Cocktail Party*. doi: 10.1007/978-3-319-51662-2.

Shinn-Cunningham, B. and Ihlefeld, A. (2004) ‘Selective and divided attention : Extracting information from simultaneous sound sources’, in *ICAD 04-Tenth Meeting of the International Conference on Auditory Display*. Sydney, Australia.

Skoglund, T. H. *et al.* (2020) ‘Big Five personality profiles in the Norwegian Special Operations Forces’, *Frontiers in Psychology*, 11(May), pp. 1–11. doi: 10.3389/fpsyg.2020.00747.

Sluiter, J. K. and Frings-Dresen, M. H. W. (2007) ‘What do we know about ageing at work? Evidence-based fitness for duty and health in fire fighters’, *Ergonomics*, 50(11), pp. 1897–1913. doi: 10.1080/00140130701676005.

Smalt, C. J. *et al.* (2020) ‘The Effect of Hearing-Protection Devices on Auditory Situational Awareness and Listening Effort’, *Ear and Hearing*, (Berger 2003), pp. 82–94. doi: 10.1097/AUD.0000000000000733.

Smith, S. B. and Cone, B. (2021) ‘Efferent unmasking of speech-in-noise encoding?’, *International Journal of Audiology*, 60(9), pp. 677–686. doi: 10.1080/14992027.2020.1862425.

Smits, Cas *et al.* (2004) ‘Development and validation of an automatic speech-in-noise screening test by telephone’, *International Journal of Audiology*, 43(1), pp. 15–28. doi: 10.1080/14992020400050004.

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’, *The Journal of the Acoustical Society of America*, 91, p. 421. doi: 10.1121/1.402729.

Soli, S. D., Giguère, C., *et al.* (2018) ‘Evidence-Based Occupational Hearing Screening I: Modeling the Effects of Real-World Noise Environments on the Likelihood of Effective Speech Communication’, *Ear and hearing*, 39(3), pp. 436–448. doi: 10.1097/AUD.0000000000000547.

Soli, S. D., Amano-Kusumoto, A., *et al.* (2018) ‘Evidence-based occupational hearing screening II: validation of a screening methodology using measures of functional hearing ability’, *International Journal of Audiology*. Taylor & Francis, 57(5), pp. 323–334. doi: 10.1080/14992027.2017.1411623.

Soli, S. D., Roeser, R. and Vaillancourt, V. (2019) ‘Changing the audiological mindset about fitness for duty assessments’, *Audiology Today*, 31(2), pp. 49-59.

Soli, S. D. and Wong, L. L. N. (2008a) ‘Assessment of speech intelligibility in noise with the Hearing in Noise Test’, *International Journal of Audiology*, 47(6), pp. 356–361. doi: 10.1080/14992020801895136.

Soli, S. D. and Wong, L. L. N. (2008b) ‘Assessment of speech intelligibility in noise with the Hearing in Noise Test’, *International Journal of Audiology*, 47(6), pp. 356–361. doi: 10.1080/14992020801895136.

Soto, C. J. *et al.* (2011) ‘Age differences in personality traits From 10 to 65: Big Five domains and facets in a large cross-sectional sample’, *Journal of Personality and Social Psychology*, 100(2), pp. 330–348. doi: 10.1037/a0021717.

Stacey, D. G. and Kurunathan, T. M. (2015) ‘Noncognitive indicators as critical predictors of students’ performance in dental school’, *Journal of Dental Education*, 79(12), pp. 1402–1410. doi: 10.1002/j.0022-0337.2015.79.12.tb06039.x.

Stenbäck, V., Hällgren, M. and Larsby, B. (2016) ‘Executive functions and working memory capacity in speech communication under adverse conditions’, *Speech, Language and Hearing*, 19(4), pp. 218–226. doi: 10.1080/2050571X.2016.1196034.

Stenfelt, S. *et al.* (2011) ‘E-health technologies for adult hearing screening’, *Audiology Research*, 1(e14), pp. 55-57. doi: 10.4081/audiores.2011.e14.

Stevens, G. *et al.* (2013) ‘Global and regional hearing impairment prevalence: An analysis of 42 studies in 29 countries’, *European Journal of Public Health*, 23(1), pp. 146–152. doi: 10.1093/eurpub/ckr176.

Stewart, G. L. (1999) ‘Trait bandwidth and stages of performance: assessing differential effects for concientiousness and its subtraits’ *Journal of Applied Psychology*, 84(6), pp. 959–968. https://doi.org/10.1037/0021-9010.84.6.959

Strasburger, H. (2001) ‘Converting between measures of slope of the psychometric function’, *Perception and Psychophysics*, 63(8), pp. 1348–1355.

St. Onge, P. *et al.* (2011) ‘Marine Corps Breacher training study: Auditory and vestibular findings’, *U.S.Army Medical Department journal*, pp. 97-107.

Sussman, E. S. (2017) ‘Auditory scene analysis: an attention perspective’, *Journal of Speech*

*Language and Hearing Research*, 60(10), p. 2989. doi: 10.1044/2017_JSLHR-H-17-0041.

Swaminathan, J. *et al.* (2016) 'Role of binaural temporal fine structure and envelope cues in cocktail-party listening', *Journal of Neuroscience*, 36(31), pp. 8250–8257. doi: 10.1523/JNEUROSCI.4421-15.2016.

Taylor, T.H. (2010) 'Ceiling effect', in N.J Salkind (ed.) *Encyclopedia of Research Design* Los Angeles, CA: Sage Reference USA, PP.132-134.

Terracciano, A. *et al.* (2005) 'Hierarchical linear modeling analyses of NEO-PI-R Scales in the Baltimore longitudinal study of aging', *Psychology and Aging*, 20(3), pp. 493–506. doi: 10.1037/0882-7974.20.3.493.Hierarchical.

Tett, R. P., Jackson, D. N. and Rothstein, M. (1991) 'Personality measures as predictors of job performance: a meta-analytic review', *Personnel Psychology*, 44(4), pp. 703–742. doi: 10.1111/j.1744-6570.1991.tb00696.x.

Theodoridis, G. C. and Schoeny, Z. G. (1990) 'Procedure learning effects in speech perception tests', *International Journal of Audiology*, 29(4), pp. 228–239. doi: 10.3109/00206099009072854.

Theunissen, M., Swanepoel, D. W. and Hanekom, J. (2009) 'Sentence recognition in noise: Variables in compilation and interpretation of tests', *International Journal of Audiology*, 48(11), pp. 743–757. doi: 10.3109/14992020903082088.

Thoresen, C. J. *et al.* (2004) 'The Big Five personality traits and individual job performance growth trajectories in maintenance and transitional job stages', *Journal of Applied Psychology*, 89(5), pp. 835–853. doi: 10.1037/0021-9010.89.5.835.

Timothy Church, A. *et al.* (2008) 'Prediction and cross-situational consistency of daily behavior across cultures: Testing trait and cultural psychology perspectives', *Journal of Research in Personality*, 42(5), pp. 1199–1215. doi: 10.1016/j.jrp.2008.03.007.

Treutwein, B. (1995) 'Adaptive psychophysical procedures', *Vision Research*, 35(17), pp. 2503–2522. doi: 10.1016/0042-6989(95)00016-X.

Treutwein, B. and Strasburger, H. (1999) 'Fitting the psychometric function', *Perception and Psychophysics*, 61(1), pp. 87–106. doi: 10.3758/BF03211951.

Trevethan, R. (2017) 'Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice', *Frontiers in Public Health*, 5(November), pp. 1–7. doi: 10.3389/fpubh.2017.00307.

Trine, A. and Monson, B. B. (2020) ‘Extended high frequencies provide both spectral and temporal information to improve speech-in-speech recognition’, *Trends in Hearing*, 24. doi: 10.1177/2331216520980299.

Tufts, J. B. *et al.* (2018) ‘Development and implementation of a military auditory fitness-for-duty test battery’, *The Journal of the Acoustical Society of America*, 143(3), pp. 1780–1780. doi: 10.1121/1.5035825.

Tufts, J. B., Vasil, K. a and Briggs, S. (2009) ‘Auditory fitness for duty: a review.’, *Journal of the American Academy of Audiology*, 20(9), pp. 539–557. doi: 10.3766/jaaa.20.9.3.

Vaez, N., Desgualdo-Pereira, L. and Paglialonga, A. (2014) ‘Development of a test of suprathreshold acuity in noise in Brazilian Portuguese: A new method for hearing screening and surveillance’, *BioMed Research International*, 2014, pp. 1–9. doi: 10.1155/2014/652838.

Vaillancourt, V. *et al.* (2011) ‘Evaluation of Auditory Functions for Royal Canadian Mounted Police Officers’, *Journal of the American Academy of Audiology*, 22(6), pp. 313–331. doi: 10.3766/jaaa.22.6.2.

Van Engen, K. J. (2012) ‘Speech-in-speech recognition: A training study’, *Language and Cognitive Processes*, 27(7–8), pp. 1089–1107. doi: 10.1080/01690965.2012.654644.

Vermiglio, Andrew J. *et al.* (2012) ‘The relationship between high-frequency pure-tone hearing loss, hearingin noise test (HINT) thresholds, and the articulation index’, *Journal of the American Academy of Audiology*, 23(10), pp. 779–788. doi: 10.3766/jaaa.23.10.4.

Vermiglio, A. J. *et al.* (2020) ‘The effect of stimulus audibility on the relationship between pure-tone average and speech recognition in noise ability’, *Journal of the American Academy of Audiology*, 31(03), pp. 224–232.

Vieira, S. and Corrente, J. E. (2011) ‘Statistical methods for assessing agreement between double readings of clinical measurements’, *Journal of Applied Oral Science*, 19(5), pp. 488–492. doi: 10.1590/S1678-77572011000500009.

Vlaming, M. S. M. G. *et al.* (2014) ‘Automated screening for high-frequency hearing loss’, *Ear and Hearing*, 35(6), pp. 667–679. doi: 10.1097/AUD.0000000000000073.

Vos, T. *et al.* (2017) ‘Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: A systematic analysis for the Global Burden of Disease Study 2016’, *The Lancet*, 390(10100), pp. 1211–1259. doi: 10.1016/S0140-6736(17)32154-2.

Wagener, K. C. and Brand, T. (2005a) ‘Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: Influence of measurement procedure and masking parameters’, *International Journal of Audiology*, 44(3), pp. 144–156. doi: 10.1080/14992020500057517.

Wagener, K. C. and Brand, T. (2005b) ‘Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: Influence of measurement procedure and masking parameters’, *International Journal of Audiology*, 44(3), pp. 144–156. doi: 10.1080/14992020500057517.

Wagner-Hartl, V., Grossi, N. R. and Kallus, K. W. (2018) ‘Impact of age and hearing impairment on work performance during long working hours’, *International Journal of Environmental Research and Public Health*, 15(1), pp. 1–13. doi: 10.3390/ijerph15010098.

Wang, H. *et al.* (2012) ‘Are the effects of conscientiousness on contextual and innovative performance context specific? Organizational culture as a moderator’, *International Journal of Human Resource Management*, 23(1), pp. 174–189. doi: 10.1080/09585192.2011.561246.

Wang, Q., Liao, Y. and Burns, G. N. (2021) ‘General, work-specific, and work-role conscientiousness measures in predicting work criteria: A comparative Perspective’, *Applied Psychology*, 70(1), pp. 358–383. doi: 10.1111/apps.12234.

Warzybok, A., Zokoll, M. A. and Kollmeier, B. (2016) ‘Development and evaluation of the Russian digit triplet test’, *Acta Acustica united with Acustica*, 102(4), pp. 714–724. doi: 10.3813/AAA.918988.

Wasiuk, P. A. *et al.* (2022) ‘Factors predicting speech-in-speech recognition : Short-term audibility , talker sex , and listener factors’, *Acoustical Society of America*, 3010. doi: 10.1121/10.0015228.

Watson, P. F. and Petrie, A. (2010) ‘Method agreement analysis: A review of correct methodology’, *Theriogenology*, 73(9), pp. 1167–1179. doi: 10.1016/j.theriogenology.2010.01.003.

Wells, C. S. and Sireci, S. G. (2020) ‘Evaluating random and systematic error in student growth percentiles’, *Applied Measurement in Education*, 33(4), pp. 349–361. doi: 10.1080/08957347.2020.1789139.

Whitton, J. P. *et al.* (2017) ‘Audiomotor perceptual training enhances speech intelligibility in background noise’, *Current Biology*, 27(21), pp. 3237-3247.e6. doi: 10.1016/j.cub.2017.09.014.

Wichmann, F. A. and Hill, N. J. (2001) ‘Wichmann psychometric and Hill I: fitting, sampling, and goodness–of–fit’, *Perception & Psychophysics*, 63(8), pp. 1293--1313. doi: 10.3758/BF03194544.

Wild, C. J. *et al.* (2012) 'Effortful listening: The processing of degraded speech depends critically on attention.', *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 32(40), pp. 14010–21. doi: 10.1523/JNEUROSCI.1528-12.2012.

Willberg, T. *et al.* (2020) 'The long-term learning effect related to the repeated use of the Finnish matrix sentence test and the Finnish digit triplet test', *International Journal of Audiology*. Taylor & Francis, 59(10), pp. 753–762. doi: 10.1080/14992027.2020.1753893.

Williams-Sanchez, V. *et al.* (2014) 'Validation of a screening test of auditory function using the telephone', *Journal of the American Academy of Audiology*, 25(10), pp. 937–951. doi: 10.3766/jaaa.25.10.3.

Wilson, B. S. *et al.* (2017) 'Global hearing health care: New findings and perspectives', *The Lancet*, 390(10111), pp. 2503–2515. doi: 10.1016/S0140-6736(17)31073-5.

Wilson, R. H. (2011) 'Clinical experience with the words-in-noise test on 3430 veterans: Comparisons with pure-tone thresholds and word recognition in quiet.', *Journal of the American Academy of Audiology*, 22(7), pp. 405–23. doi: 10.3766/jaaa.22.7.3.

Wong, L. L. N., Ng, E. H. N. and Soli, S. D. (2012) 'Characterization of speech understanding in various types of noise', *The Journal of the Acoustical Society of America*, 132(4), pp. 2642–2651. doi: 10.1121/1.4751538.

Xi, X. *et al.* (2012) 'Development of a corpus of Mandarin sentences in babble with homogeneity optimized via psychometric evaluation', *International Journal of Audiology*, 51(5), pp. 399–404. doi: 10.3109/14992027.2011.642011.

Xia, J. *et al.* (2018) 'Effects of reverberation and noise on speech intelligibility in normal-hearing and aided hearing-impaired listeners', *The Journal of the Acoustical Society of America*, 143(3), pp. 1523–1533. doi: 10.1121/1.5026788.

Yalkawi, H. (2022) Telephone conversation with Iman Rawas, 4 December.

Yund, E. W. and Woods, D. L. (2010) 'Content and procedural learning in repeated sentence tests of speech perception', *Ear and Hearing*, 31(6), pp. 769–778. doi: 10.1097/AUD.0b013e3181e68e4a.

Zaballos, M. T. P. *et al.* (2015) 'Effects of long-term speech-in-noise training in air traffic controllers and high frequency suppression. A control group study', *Journal of International Advanced Otology*, 11(3), pp. 212–217. doi: 10.5152/iao.2015.1745.

Zendel, B. R. *et al.* (2019) ‘Musical training improves the ability to understand speech-in-noise in older adults’, *Neurobiology of Aging*, 81, pp. 102–115. doi: 10.1016/j.neurobiolaging.2019.05.015.

Zhang, P. *et al.* (2019) ‘Evaluating the performance of the staircase and quick change detection methods in measuring perceptual learning’, *Journal of Vision*, 19(7), pp. 1–25. doi: 10.1167/19.7.14.

Zhang, Y. *et al.* (2017) ‘Personality traits and perception of Müller-Lyer illusion in male Chinese military soldiers and university students’, *Translational Neuroscience*, 8(1), pp. 15–20. doi: 10.1515/tnsci-2017-0004.

Zou, K. H., O’Malley, A. J. and Mauri, L. (2007) ‘Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models’, *Circulation*, 115(5), pp. 654–657. doi: 10.1161/CIRCULATIONAHA.105.594929.

Zumbo, B. D. (2016) ‘Standard-setting methodology: Establishing performance standards and setting cut-scores to assist score interpretation’, *Applied physiology, nutrition, and metabolism = Physiologie appliquee, nutrition et metabolisme*, 41(6), pp. S74–S82. doi: 10.1139/apnm-2015-0522.